Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions

REBECCA ADAIMI, University of Texas at Austin, USA
HOWARD YONG, University of Texas at Austin, USA
EDISON THOMAZ, University of Texas at Austin, USA

Conversational assistants in the form of stand-alone devices such as Amazon Echo and Google Home have become popular and embraced by millions of people. By serving as a natural interface to services ranging from home automation to media players, conversational assistants help people perform many tasks with ease, such as setting timers, playing music and managing to-do lists. While these systems offer useful capabilities, they are largely passive and unaware of the human behavioral context in which they are used. In this work, we explore how off-the-shelf conversational assistants can be enhanced with acoustic-based human activity recognition by leveraging the short interval after a voice command is given to the device. Since always-on audio recording can pose privacy concerns, our method is unique in that it does not require capturing and analyzing any audio other than the speech-based interactions between people and their conversational assistants. In particular, we leverage background environmental sounds present in these short duration voice-based interactions to recognize activities of daily living. We conducted a study with 14 participants in 3 different locations in their own homes. We showed that our method can recognize 19 different activities of daily living with average precision of 84.85% and average recall of 85.67% in a leave-one-participant-out performance evaluation with 30-second audio clips bound by the voice interactions.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Smart Environment, Smart Speaker, Voice Assistants, Conversational Assistants, Google Home, Environmental Sounds, Activities of Daily Living, Audio Processing, Deep Learning, Human Activity Recognition

ACM Reference Format:
Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 2 (March 2021), 24 pages. https://doi.org/10.1145/3448090

1 INTRODUCTION

Conversational assistants like Amazon Echo and Google Home [33] have recently become mainstream. These systems can perform a number of tasks on-command such as control smart home appliances, play music, and answer questions, through a voice-based interface, e.g., "Alexa, what day is today?", "Ok Google, turn on the living room lights". Recent survey reports have shown that 54% of the U.S. population use voice-command technologies, whether on smartphones, smart speakers, and other devices, and 24% reported using voice assistants daily [31]. By 2025, it is estimated that 75% of U.S. households, representing over 100 million households, will have smart speakers [20].

Authors' addresses: Rebecca Adaimi, University of Texas at Austin, 2501 Speedway, Austin, Texas, 78712, USA, rebecca.adaimi@utexas.edu; Howard Yong, University of Texas at Austin, 2501 Speedway, Austin, Texas, 78712, USA, howardyong@utexas.edu; Edison Thomaz, University of Texas at Austin, 2501 Speedway, Austin, Texas, 78712, USA, ethomaz@utexas.edu.

2474-9567/2021/3-ART2 $15.00
https://doi.org/10.1145/3448090

While voice-based assistants have proved useful to millions of people, they are still limited in numerous ways. For instance, today's assistants are completely passive, simply responding to commands. If they could become aware of people's behavioral and environmental context, it would be possible to integrate them into the flow of everyday life more seamlessly, enabling much richer applications. As an example, imagine an assistant helping a person cook a new dish for the first time. It could observe and make sure all recipe steps are being followed in the right order; likewise, it could keep track of cooking times, set cooking temperatures, and turn appliances on and off as needed. Or consider an assistant recognizing when a person is having trouble using a tool, performing a repair or building furniture, and proactively looking up and displaying instructions and how-to videos. Lastly, imagine an assistant picking up early signs of cognitive impairment in older adults by detecting subtle changes in the way they ask questions and perform household tasks.

Motivated by these types of scenarios and applications, this paper explores how off-the-shelf conversational assistants can be enhanced with acoustic-based human context awareness to recognize activities of daily living (ADL). Since continuous audio recording in a home setting is largely undesirable due to privacy concerns, we describe a method that does not require the capture of any audio other than what is already recorded as part of the voice-based interactions. The method hinges on the observation that these acoustic interactions are often accompanied by background environmental sounds that can be a rich source of context. For example, when people make a request such as "Alexa, what is the weather?", the audio of the request, which is captured by the assistant, often contains contextual information in the background, such as conversations taking place, kids playing, babies crying, TV sounds, cooking sounds, appliances running, toilets flushing and more.

Technically, this approach encounters important data-related challenges. Crucially, previous work in acoustic-based activity recognition has typically relied on unrestricted and continuous audio capture, which results in large amounts of data that can be mined for activities and context [24, 26]. Here, we are constrained by the audio data present in the voice assistant interactions, which limits the quantity and type of data captured. Additionally, we investigate how background acoustics can be analyzed in the presence or absence of speech as well as when using the whole acoustic interaction vs. only the *mid-interaction* segment. The mid-interaction segment refers to the pause between the user query and the voice assistant's response and is typically no more than 3 seconds. Given its short duration, it is particularly challenging to recognize meaningful background contextual information from it. The paper makes the following contributions:

- A comprehensive quantitative evaluation of our acoustic-based recognition framework in people's own homes. We conducted a study with 14 participants in 3 different locations in their own homes. We show that our method can recognize 19 different activities of daily living with average precision of 84.85% and average recall of 85.67% in a leave-one-participant-out performance evaluation with 30-second clips.
- Automatic inference of the voice assistant's location in the home based on the recognition of activities performed with average 96.81% precision and 95.24% recall.
- An annotated dataset of audio interactions with a conversational assistant with background sounds of 19 activities. We also make available the source code of our processing pipeline to encourage others to replicate and build upon our work.

2 RELATED WORK

2.1 Conversational Assistants

With conversational assistants becoming ubiquitous in people's lives, several researchers have taken this opportunity to better understand the general use of such devices in the home environment and how people perceive them. Bentley *et al.* studied in depth the long-term use of smart speaker assistants in order to better understand exactly how users are using these devices over time [5]. Lahoual *et al.* investigated the efficiency of vocal interactions with voice assistants (VA) by presenting a qualitative user-centered study [23]. They showed that, despite the use of VAs leading to additional supervision, verification, diagnosis and problem-solving activities that cause an interruption in domestic activity, users continued to use and accept the system. Similarly, Kiseleva *et al.* focused on understanding user satisfaction with voice assistants over a range of typical scenarios of use, such as controlling a device or web search [21]. They found that user satisfaction varies across different scenarios and sometimes depends on either task completion or perceived effort spent on the task. Looking at how users personify such devices, Pradhan *et al.* studied how older adults perceive social interactions with the voice assistant "Alexa" and ontologically categorize the agent as "human-like" or "object-like" [35]. Going beyond current voice assistants, Tabassum *et al.* went even further by investigating the next generation of voice assistants, "always listening voice assistants", that could passively listen to people's conversations and proactively provide assistance. [41]. They explored the potential services people anticipate from such a device and how they feel about sharing their data for these purposes. Other researchers have studied the use of voice assistants by focusing on particular sub-populations, such as multi-user households [33], children and parents [12, 15], people with disabilities [36], older adults [35], and under-served users in low-income regions [38].

Despite substantial prior work aimed at understanding usage of voice assistants in daily life across different populations, no previous efforts attempted to take advantage of the assistant's interaction process towards activity recognition, which is the focus of this work.

2.2 Activities of Daily Living

Detecting human activities in the home has been a motivation of the field of human activity recognition for many years. Particular emphasis has been given to so-called Activities of Daily Living (ADL), which are essential activities a person needs to undertake daily in order to maintain independence [45]. The automatic identification of ADLs is an important and useful task, especially in elderly and patient care [10].

Many previous efforts have relied on commodity sensors such as accelerometers and gyroscopes found in smartphones, smart watches, and wearables for activity recognition. These sensors produce motion-based features that are useful to detect many relevant health behaviors such as eating, walking, running, etc. [8, 9, 29], and identifying life patterns [48]. However, ADLs typically include complex activities that common single sensing approaches often fail to capture [42]. As a result, researchers have turned to richer sensing modalities and approaches such as sound, images, and multimodal methods.

In the home, a common approach has been to use environmental sensors, e.g., camera-based sensors [3, 7, 25], thermal sensors [19], or infrared sensors [44, 49] for motion and location detection. These types of sensors are able to detect appliance usage, measure water flow, and perform indoor localization. Several researchers have also experimented with image-based sensors installed on walls and ceilings, to locate residents and recognize their activities [3, 7, 25]. However, image-based methods raise privacy concerns for in-home monitoring. To address these concerns, Hevesi *et al.* instrumented a household with several infrared sensor arrays to monitor various daily household activities while providing a less privacy-invasive ADL monitoring system [19]. In the development of sensing environments, several works have explored other monitoring technologies, such as pressure, floor, and radar sensors, that make it less invasive to the user [4, 14, 27]. Overall, instrumenting homes with specific sensors tends to be both undesirable and impractical in the long-term.

2.3 Acoustic Activity Recognition

A promising way of monitoring activities in the home is by sensing sounds, specifically environmental sounds. A household environment includes several items that are commonly used for performing daily tasks, like the electrical toothbrush, the shaver, the washing machine, the sink, the stove, etc. Most of those appliances and fixtures create or disperse sounds while being used, which can in turn indicate the activity being performed. Dimitrov *et al.* investigated environmental sounds for touch-free audio-based device recognition in a home environment [11]. A room was instrumented with a microphone, and a device recognition framework was implemented to recognize different devices from their characteristic sounds. Vacher *et al.* presented the SWEET-HOME project, whose goal was to collect a multimodal dataset for developing an audio-based smart home environment [46]. The home was instrumented with several microphones to capture different sounds from different parts of the house. Tremblay *et al.* developed an activity recognition system based on environmental sounds that is able to detect errors related to cognitive impairment [43]. Odashima *et al.* performed sound-based activity recognition via unsupervised clustering of human-based and machine-based sounds [32]. Laput *et al.* utilized public sound effect libraries to develop a plug-and-play real-time sound-based activity recognition system [24]. They analyzed their approach across several devices, such as a smartwatch, smartphone, and laptop and provided insights into the feasibility of sound-based activity recognition. Incorporating a mixed process of audio augmentation and combining online sound effect libraries with the Audio Set data, they presented a system for audio context classification. Another similar work is presented by Liang *et al.* where an audio-based framework is proposed that combines large-scale on-line YouTube video soundtracks with oversampling to train activity recognition models [26]. They tested their model on 15 activity classes collected in the wild. However, these previous works are mainly situated in a setting where acoustic data is continuously captured. Thus, this raises several privacy concerns.

In our work, we aimed to study the feasibility of acoustic activity recognition when constrained by when, how, and how long data can be captured. Different from previous work, we therefore capture audio data under the constraints of voice assistant interactions. We show that, using limited amount of data collected in people's own natural environment, we are still able to recognize activities occurring in the background during an interaction with the voice assistants. Moreover, we show that, by applying transfer learning using AudioSet without additional data augmentation or oversampling, we are able to train a classifier with limited real-world data.

3 MOTIVATION AND APPROACH

Due to the natural form of interaction made possible by voice-based conversational assistants, conversational interfaces are easy to use and have the potential to collect and process a large number of commands over the course of weeks and months. Consequently, much could be learned about the individuals who own and use such devices, such as their preferences, health condition, demographics, family composition, the environment in which they live, and activities of daily living (ADLs). In particular, automating the tracking of ADLs is essential in elderly and patient care. Instead of resorting to nursing homes or paid home care, imagine being able to remotely monitor routine ADLs of an elderly relative living alone without using wearables or instrumenting the household with environmental sensors. By utilizing voice assistants, which have been adopted by millions of people, it is possible to consider monitoring loved ones seamlessly and comfortably while maintaining them in their homes.

When approaching this problem, and considering alternatives for data capture, a key design goal was to ensure that individuals maintain their ability to use the conversational assistants in its full capacity. Initially, we investigated utilizing a separate device that independently detects wake-words. Porcheron *et al.* developed such a device called Conditional Voice Recorder [34]. We also explored integrating the Google Assistant into a separate device, such as a Raspberry Pi, using the Google Assistant SDK [17]. However, for both approaches, relying on a separate device to detect the wake-word proved challenging mainly in synchronizing it with the Google Assistant. Thus, in order to gain better reliability in capturing a user's interaction with a smart speaker, we designed a hardware add-on that is linked to the Google Home's activation by the voice command. Section 4 provides a detailed explanation of our device implementation, and section 8.6 discusses privacy considerations underlying our approach.

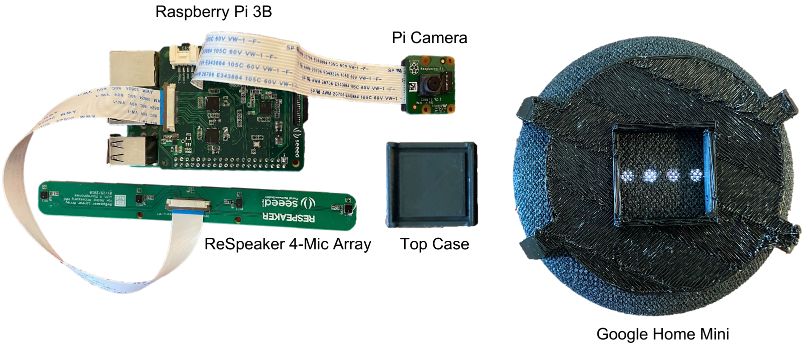


Fig. 1. Hardware setup of our *add-on* device.

4 AUDIO CAPTURE ACCESSORY FOR CONVERSATIONAL ASSISTANT

To capture voice-based interactions with the assistant, we developed an audio recording *add-on accessory* that does not interfere with the assistant's operation and functionalities. The accessory, which may also be helpful in studying accidental misactivations of smart speakers [13], will be made open source and available to the community. The next sections describe the accessory's hardware and software implementations.

4.1 Hardware Implementation

In terms of the hardware implementation, our *add-on* device was developed using a Raspberry Pi, a ReSpeaker 4-mic array [40], and a camera module [28] (Figure 1). Our device records and collects its own data via the 4-mic array attached to the Raspberry Pi, allowing greater flexibility and control over the data collection process. This also allowed for easy programmatic access when collecting, storing, and manipulating the acoustic data. Contrary to previous audio-based ADL-related work, our system does not passively record and store audio data, but instead is controlled by the user's interaction with the Google Home. In other words, similar to how Google Home is always listening to the voice command and only records the interaction, our system is always looking for the Google Home's top LED lights to turn on, indicating the voice assistant heard the voice command "Hey Google" or "Ok Google" and is awake. Once this occurs, our system records the interaction and saves the audio data locally. Refer to Section 4.2 for more details on the software implementation of this functionality. In order to sense the LED lights, we utilized a camera module and securely placed it on top of the Google Home via a 3D-printed mounting structure.

4.2 Software Implementation

To reiterate, our *add-on* system is designed to activate when the smart speaker is activated by the voice command, which is indicated by the LED lights. To that end, the Raspberry Pi is programmed to continuously monitor the LED light activity using the camera module and averaging the pixel brightness of the image stream. For our purposes, we did not require a high frame rate or high quality image since the camera was acting as a light sensor to detect when the top LED lights of the Google Home device turn on. Thus, after several empirical tests, a frame rate of around 2 fps and image resolution of 128×80 proved suitable for this application. This also helps reduce the overall processing load on the Raspberry Pi. Meanwhile, the 4-mic array is also continuously listening, but without saving the recorded audio. After several designs, we noticed a delay that occurs between detecting

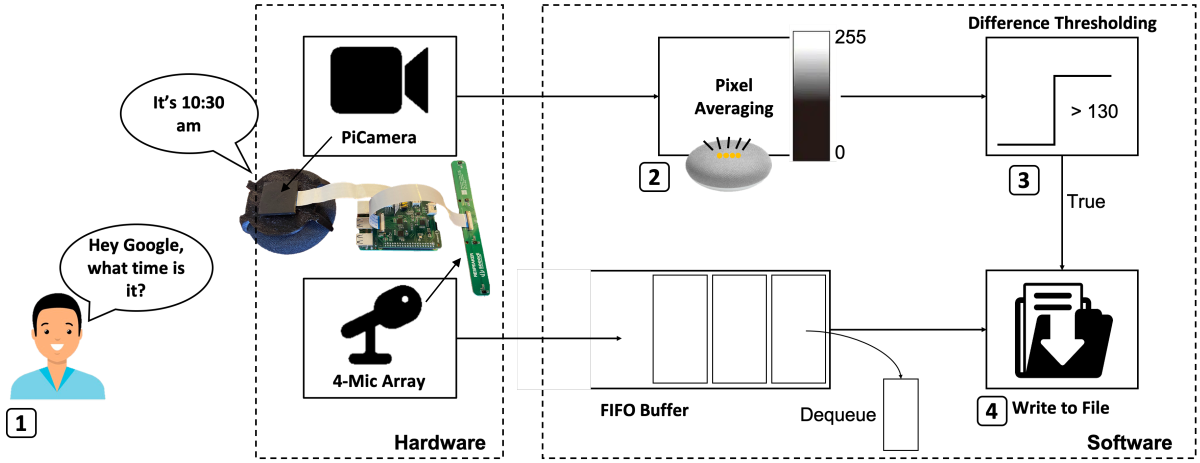


Fig. 2. Logic flow implementation of add-on system. Once the system is turned on, the camera continuously checks if the LED lights turn on while the microphone captures audio stored in a fixed size FIFO buffer. The system is triggered when first (1) the user interacts with the voice assistant which (2) triggers the LED lights to turn on. (3) The pixel average computed from the image captured by the camera will in turn exceed the threshold which (4) triggers the system to save the audio data stored in the buffer.

the LED lights and starting the recording process. Since interactions are typically very fast, in some cases the user's interaction with the voice assistant is not captured. Thus, in order to mitigate this issue, we designed our system to continuously listen and store 30 seconds of audio in a First-In-First-Out (FIFO) buffer. In this way, when the voice assistant receives an inquiry, it activates the LED lights which, captured by the camera, triggers a new thread process that writes the audio clip stored in the FIFO buffer to a wav file. With this functionality, we are given more control over the size of the audio clip recorded as well as the amount of information captured before and after the interaction, which can also be helpful in capturing more context information. Taking into account that we are constrained by the voice assistant's interaction duration and in order to investigate how performance changes with longer vs. shorter sound clips, we set the buffer size to 30 seconds with a wait time of 15 seconds after the LED light turns on. This allows us to capture around 10 seconds of audio before and after the interaction. The audio is sampled at 22KHz, which was the maximum firmware sample rate allowed by the 4-mic array. Moreover, analyzing the frequency domain of environmental audio sounds, a sample rate of 22KHz was observed to be good enough to capture the difference of sounds. It is important to note that despite our system continuously listening, it overwrites the audio information every 30 seconds and only stores the data once triggered by the Google Home. The system workflow is depicted in Figure 2.

5 ACOUSTIC ACTIVITY RECOGNITION

As described in the previous section, for every interaction with the voice assistant, a 30-second audio clip is recorded and stored locally. This audio clip includes the question-answer interaction as well as any background sounds captured during the interaction. These sounds serve as the underlying clues to infer a person's activity and one's surroundings. In this section, we present our analysis of the data in terms of acoustic feature extraction and data preprocessing that help recognize the different activities.

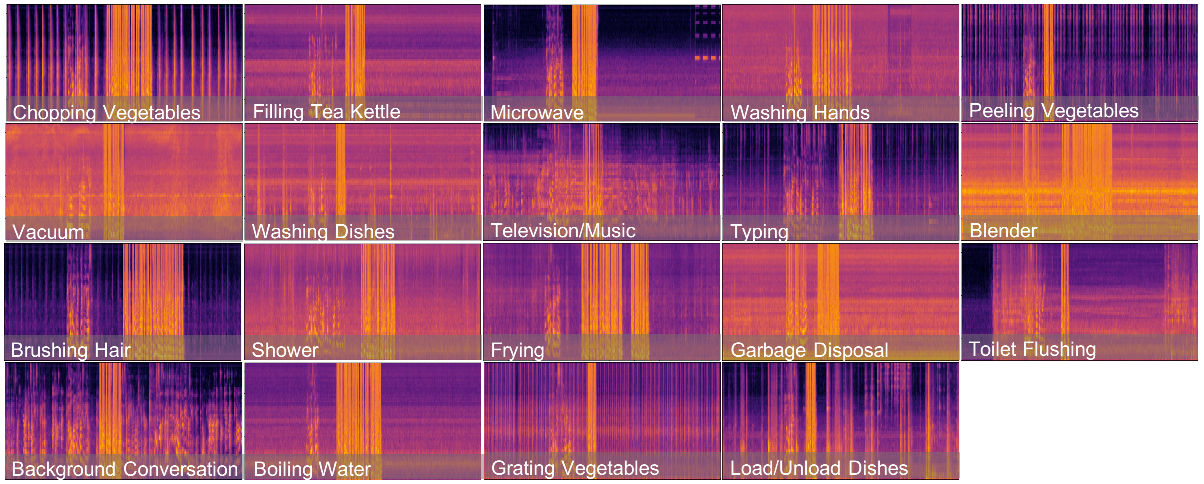


Fig. 3. Example Log Mel spectrograms of 30s audio clips for 19 activity classes. The audio clips included the interaction with Home Assistant.

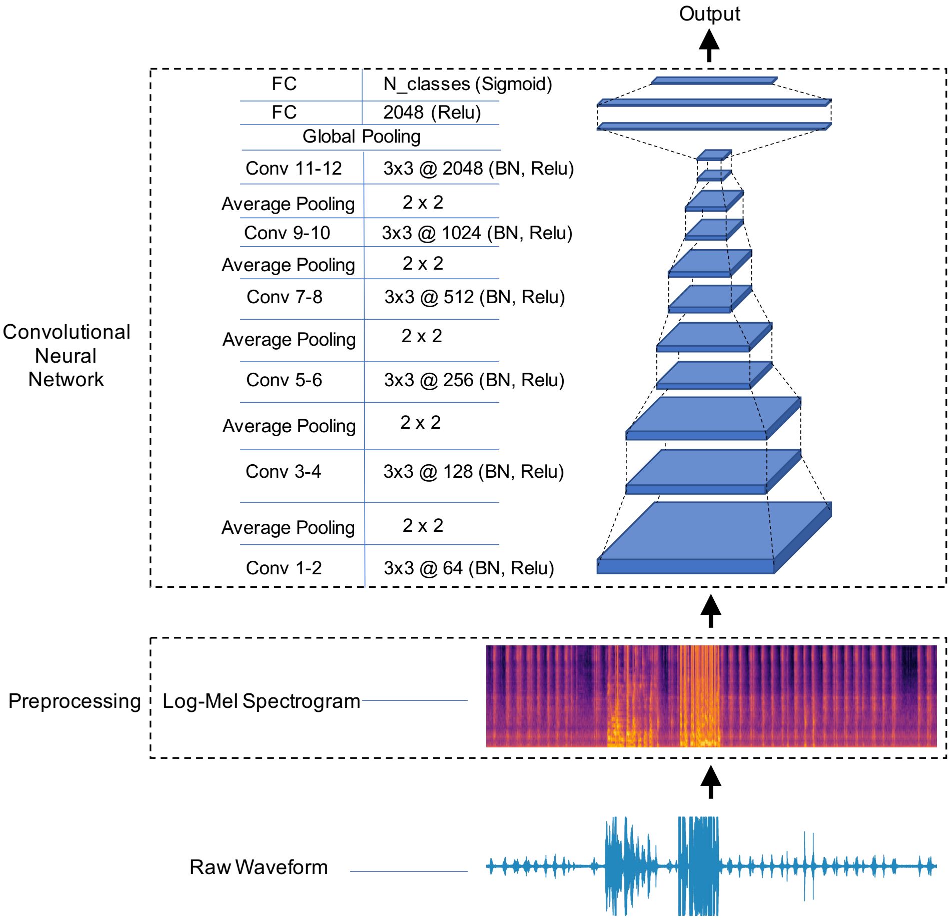
5.1 Audio Data Preprocessing

Log-mel spectrograms have proved very powerful in audio classification when used as input to deep learning models [18]. Thus, we extracted the log-mel spectrograms of our audio clips by computing the short-time Fourier transform (STFT) for each segment using a Hanning window of 1024 samples and a hop size of 320 samples. The linear spectrogram is then converted into a 64-bin log-scaled Mel spectrogram. Example log mel spectrograms of each activity class are depicted in Figure 3. As noted previously, our system captures approximately 30 seconds of audio that includes background sounds as well as the user's interaction with the home assistant. We experimented with several preprocessing approaches that included segmenting the audio clips to shorter segments vs. keeping each 30 second audio clip as one input sample. We observed higher performance when the audio clips were kept intact and thus adopted this approach for all subsequent evaluations. We further analyzed how performance changes with varying clip size.

5.2 Recognition Framework

The manual collection of ground truth audio data from individual users can be quite laborious especially when targeting multiple sound classes. Despite our efforts to collect data from a wide range of users, we were still limited by the size of our data to be able to efficiently train a deep learning model to recognize the different environment sounds. Thus, leveraging publicly available large audio datasets such as the AudioSet database [16], we utilized pre-trained models on this dataset to extract meaningful embeddings of our data that can be used as feature inputs to a classifier. Moreover, another advantage of using AudioSet is that it provides audio samples containing simultaneously occurring acoustic events including speech and other background sounds. This is especially beneficial as our system captures background sounds in audio recordings accompanied by human and machine speech.

Convolutional Neural Networks (CNNs), inspired from VGG-like network, have been proven very effective in audio classification when applied to the log-mel spectrograms of the acoustic data [18, 22]. Experimenting with several CNN architectures pre-trained on AudioSet for our task of recognizing various environmental sounds, we build upon the CNN architecture described in Figure 4, originally trained for AudioSet tagging.



FC	N_classes (Sigmoid)
FC	2048 (Relu)
Global Pooling	
Conv 11-12	3x3 @ 2048 (BN, Relu)
Average Pooling	2 x 2
Conv 9-10	3x3 @ 1024 (BN, Relu)
Average Pooling	2 x 2
Conv 7-8	3x3 @ 512 (BN, Relu)
Average Pooling	2 x 2
Conv 5-6	3x3 @ 256 (BN, Relu)
Average Pooling	2 x 2
Conv 3-4	3x3 @ 128 (BN, Relu)
Average Pooling	2 x 2
Conv 1-2	3x3 @ 64 (BN, Relu)

Fig. 4. CNN Architecture pre-trained on AudioSet. The model is used as a feature extractor and the last fully connected layer is modified and trained on our collected audio data.

The architecture contains 6 convolutional blocks with each convolutional block consisting of 2 convolutional layers (3x3 kernel) with intermediary average pooling layers. We modified this pre-trained model by removing the last fully connected layer and replacing it with our own fully connected layer, using a sigmoid activation function. Finally, we use this model as a feature extractor by freezing its parameters and only training the last fully connected layer with our collected audio data. Through experimentation and observations from previous work, a batch size of 32 and an Adam optimizer with 1e-5 learning rate is used for training for 500 epochs.

6 SEMI-NATURALISTIC DATA COLLECTION

To validate our system and our experimental design protocol and instrumentation, we first ran a formative controlled study with two participants in a designated space on campus equipped with kitchen appliances and fixtures such as microwave, sink, stove, etc.. The researchers also provided additional portable equipment needed (e.g. vacuum and blender). Participants were asked to perform a set of simple household tasks following the

Table 1. Set of activities for every location context.

Kitchen	Living Room	Bathroom
Peeling Vegetables Chopping Vegetables Grating Vegetables Frying Filling Tea Kettle Boiling Water Washing Dishes Microwave Garbage Disposal Blender	Typing (Computer) Television Background Conversation Vacuum	Washing Hands Brushing Hair Shower Toilet Flushing

Table 2. List of questions for participants to ask the Google Assistant during the user study.

Topic	Question
Weather	Hey/Ok Google, what's the weather? Hey/Ok Google, what's the weather tomorrow?
Time	Hey/Ok Google, what time is it?
Traffic	Hey/Ok Google, how's the traffic?
Information Seeking	Hey/Ok Google, how many people live in the US? Hey/Ok Google, how do you spell pineapple? Hey/Ok Google, how many calories are in a banana? Hey/Ok Google, what is 60 divided by 3? Hey/Ok Google, is walgreens still open? Hey/Ok Google, how many kilometers are in a mile? Hey/Ok Google, how do you say "Nice to meet you" in French?

same procedure designed for the semi-naturalistic study in people's homes. This helped us address issues in our experimental procedures as well as account for any potential hurdles. In particular, we determined the set of equipment needed from each participant to be able to perform the activities. We further improved the design of our 3D printed structure to ensure the camera is securely placed on top of the LED lights. This also allowed us to tune the pixel average threshold to make sure the system captures the data when an interaction occurs.

For the semi-naturalistic study, the system is required to run standalone and unattended. Thus, the Raspberry Pi is programmed to run the main software implementation autonomously on boot, i.e. users are only required to attach the system to the Google Home as shown in Figure 1 and plug it in. Additionally, in order to monitor the system for unexpected hardware issues, the device is set up to report the hardware activity status and process logs every 10 minutes via email to the primary researchers. In this way, an on-time troubleshooting can be provided to make sure everything runs smoothly and data is captured properly. We validated the smooth deployment of our system by running several additional validation studies in our own homes. The system, comprised of the Google Assistant, the *add-on* device, and the mobile hotspot, was automatically plugged in in a random location and left running unattended for more than 2 weeks during which we interacted with the voice assistant and captured data.

After validating our system and experiment design, we conducted an IRB-approved semi-naturalistic study. 15 participants of varying age (mean age 43.6 ± 13.7), profession, socioeconomic status, and gender (9 female and 6 male) were recruited through a recruiting agency. We term this study semi-naturalistic since it was conducted in people's own homes using their own tools while still following a set of instructions. Participants performed a set of simple tasks in their own homes. While performing these tasks, participants were required to interact with a Google Assistant, by asking questions such as "What is the weather?". We were unable to collect one participant's data due to network connection issues that prevented the smart speaker from working smoothly.

6.1 Activities Set

In order to evaluate our system in recognizing these types of activities, we selected three location contexts in which everyday activities occur: (1) Kitchen, (2) Living Room, and (3) Bathroom. For each context, we selected habitual activities based on the following selection criteria: 1) does the event happen frequently in that context?, 2) does it produce acoustic sounds that can be captured by a microphone?, and 3) is the event related to ADLs? In total, we selected 19 activities across 3 locations (Table 1).

6.2 Procedure

Due to social distancing requirements, we conducted a remote study. Firstly, a box with the required devices (e.g. a Google Assistant outfitted with the *add-on* device, a mobile Wifi hotspot, and study design documents) was dropped off at a participant's home. The devices work immediately by simply plugging them in for power. During the study, participants were then monitored and given further instructions by the principal investigator over a Zoom video conference call (for social distancing). Once the study was done, the participant was asked to place the devices back in the box and leave it outside the house. Researchers then picked up the box and repeated the process with the next participant.

When the study was scheduled, the participant joined a Zoom video conference call with the researchers for remote monitoring. The researchers briefly went over the consent form. As mentioned in Section 6.1, there are 3 primary locations. The study started out in the Kitchen and then moved sequentially to the Living room and then the Bathroom. For every location, the participant was asked to plug in the devices anywhere they wanted in that location, in order to capture the corresponding activities. For concrete documentation and ground truth monitoring, the Zoom call was recorded. Once the devices were plugged in, the researcher remotely made sure the system is running smoothly through the hardware process logs sent via email from the Rasberry Pi before starting the study. For every activity, a timer was displayed on the Zoom screen in order to help guide participants for when to start and stop an activity and when to interact with the Google Home. The timer was reset after every activity. At a high level, the participant started performing the activity when the timer started. At least 10 seconds in, the participant was required to interact with the Google Home by asking a question, while the activity was still being performed. After at least 30 seconds have passed, the participant was signaled to stop the activity. Each activity was performed 2 times with the participant interacting with the assistant every time. All activities, except for background conversation, involved the participant interacting with a household appliance or fixture to perform the activity. For background conversation, participants were asked to naturally start a conversation about any random topic with another household member. In case of a single-person home, participants conversed with the researchers over Zoom. Once all the activities in one location were completed, the participant was asked to move the devices to the next location. A list of possible questions encompassing various topics (weather, time, traffic, and information seeking) to ask the Google Assistant (Table 2) was provided to the participants, with the option to ask other questions if they wanted to. The questions were set based on previous related work that studied the long-term and daily usage of smart speakers [5, 15].

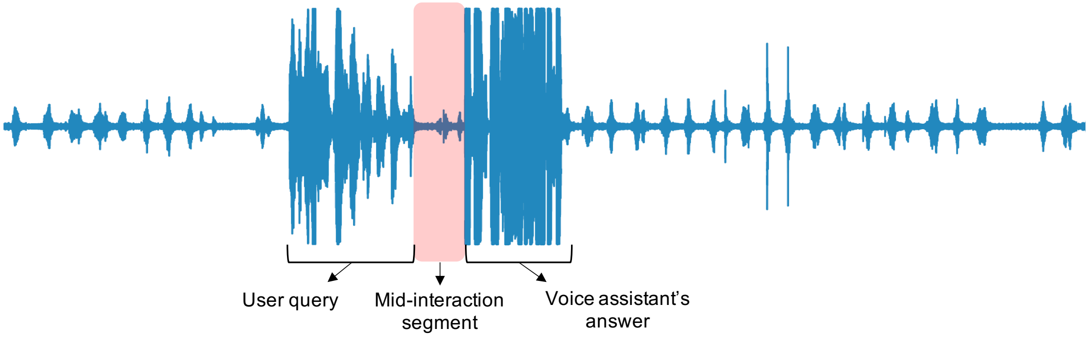


Fig. 5. Sample audio clip showing the *mid-interaction segment*.

6.3 Mid-Interaction Segment

As described earlier, the system was programmed to capture a 30-second audio clip with every interaction. An interaction with the voice assistant entails the user query (e.g. "Hey Google, what time is it?") and the assistant's answer (e.g. "It's 10:45"). We observed from the Google Assistant voice recordings that the short audio segment in-between the query and the answer does not include speech and thereby contains only background sounds. In this section, we focus on the analysis of this segment, which we define as the *mid-interaction* segment (Figure 5).

The research question we explored was: can we recognize activities when further restricted to this mid-interaction segment? We observed that the segment's length varies depending on the internet connection speed or internet interference as well as the type of question asked. On average, the duration was around 2.3 seconds. However, we encountered with some participants connectivity issues due to signal interference which caused the voice assistant to have a longer delay in its response, reaching around 14 seconds. In other cases, the segment was $<$ 1 second, typically when the question asked was straightforward such as "What time is it?". Looking at the distribution of the length of the mid-interaction segments, $\sim$ 82% were $<$ 3 seconds, $\sim$ 16% fell in the range of 3 and 7 seconds, and the remaining $\sim$ 2% going beyond 7 seconds. These longer segments can be considered as outliers and were therefore dropped from the analysis since this was caused by a connectivity slowdown attributed to our mobile hotspot. Moreover, a smart speaker will generally require a minimum bandwidth of $\sim$ 5 Mbps to work properly, which is typically available in people's homes [39]. Discarding the outliers, the mid-interaction segments were on average 2.2 seconds long (± 1.1 seconds).

For efficient training and data loading on GPUs, we explored reducing or extending the variable-sized mid-interaction clips to same lengths. To that end, several approaches were considered: (1) zero padding, (2) replication (symmetric) padding, and (3) segmenting the clips into equal-sized segments. Due to the significant difference in the clip lengths varying from $<$ 1 second to 6 seconds, zero padding can alter the data distribution. Regarding segmenting the clips into equal-sized segments, several data samples were $<$ 1 seconds long. Thus, truncating according to the length of the smallest sequence resulted in very short segments that do not capture the activity. This would inhibit the model from learning to properly distinguish the activities. In order to not disturb the data distribution as much as well as to keep a useful segment length, replication (symmetric) padding was implemented. Replication padding essentially pads the segments with its wrap. This is essential in minimizing the disruption of the data distribution compared to zero or constant padding.

It is also important to note that, in some cases and given the semi-naturalistic nature of our study, an activity was not acoustically captured in the mid-interaction segment. During the study, participants were instructed to continuously perform the activity for 30 seconds. This also allowed us to simulate a situation when people other

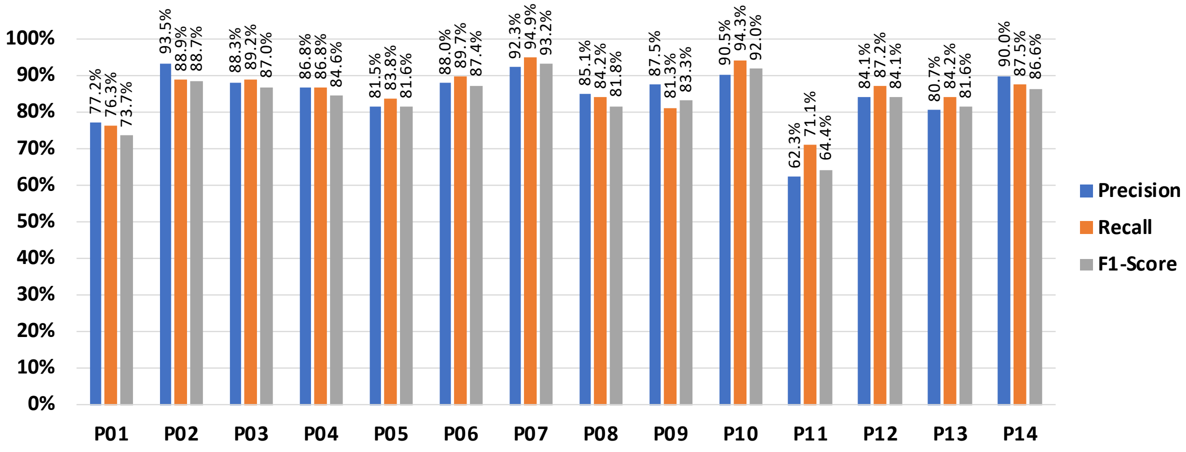


Fig. 6. Bar plot showing the performance scores of every LOPO evaluation of the location-free model using whole audio clips

than the speaker are performing the activity continuously in the background. Naturally, participants paused when asking the question and then resumed the activity right after. Some participants waited for the voice assistant to respond before resuming the activity which caused the mid-interaction segment to not capture the activity. In our dataset, we encountered such a case in only around 2% of the audio clips. For that reason, and due to the fact that in a real-world setting this is unavoidable, we did not remove such clips from our training or testing data.

7 RESULTS

In this section, we describe our results from a series of experiments wherein we investigate location-free and location-specific modelling. We conduct each experiment for both whole 30-second audio clips and mid-interaction segments. Source code for the analysis is available at https://github.com/Human-Signals-Lab/Acoustic-Activity-Recognition-Bounded-by-Conversational-Assistant-Interactions.git.

7.1 Whole Audio Clip as Input

7.1.1 Location-free Modelling. For location-free prediction, we assume the location of the device is unknown and thus train the model on all 19 classes. Given that our study was conducted in the wild, our data contains a lot of variability in terms of the living environment, the placement of the device, and the way the activities were performed. Thus, in order to effectively measure our model's performance, we apply a Leave-One-Participant-Out (LOPO) evaluation.

Using the whole 30-second audio clip, the model was able to identify the set of 19 classes with an average precision of 84.85%, average recall of 85.67%, and an average F1-score of 83.56% across the LOPO evaluations. The bar plot in Figure 6 shows the model's performance for every LOPO evaluation. For participants P1 and P11, performance was lower compared to the other LOPO evaluations. It is important to note that each participant has 2 audio clips recorded for each class which means that the model is tested on only 38 samples per participant with 2 samples per class. Looking at the mispredicted samples for participants P1 and P11, we observed that the model in some cases was not able to distinguish the different water-related activities, such as washing hands, washing dishes, shower, boiling, and filling water etc. Moreover, several other activities have similar repetitive

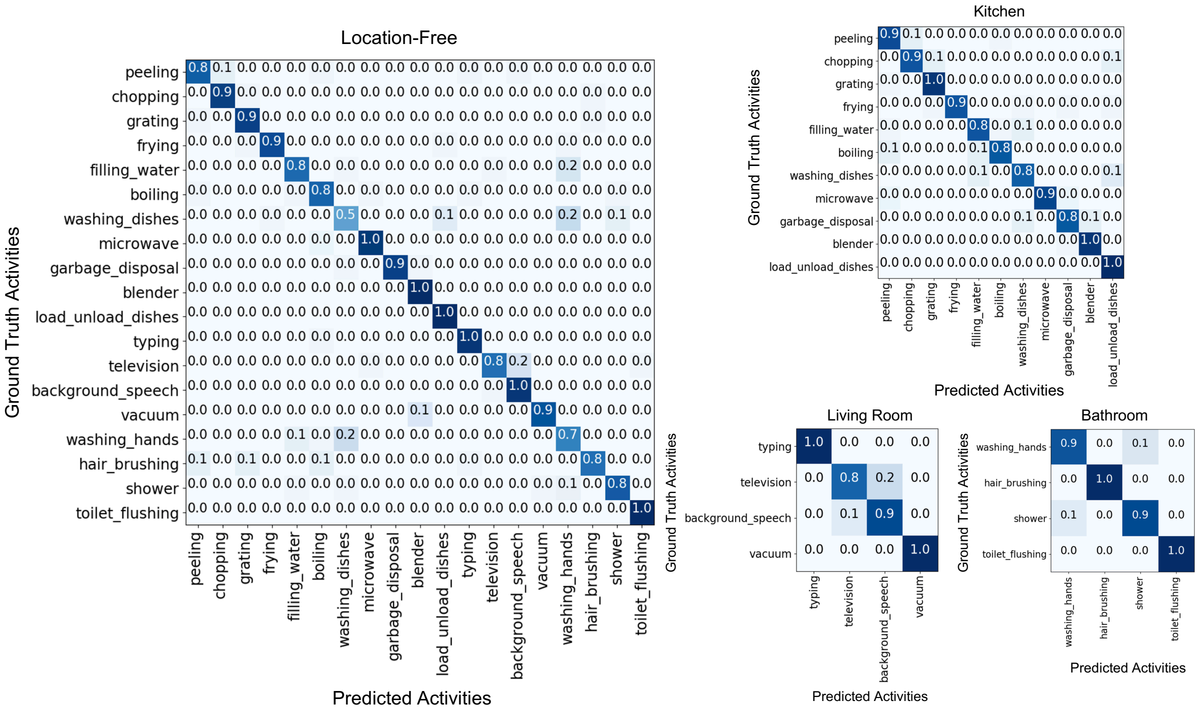


Fig. 7. Confusion matrices for activity recognition using whole 30s audio clip as input

Table 3. Performance comparison between location-free and location-specific models (kitchen, living room, bathroom)

		Whole Audio Clip			**Mid-interaction Segment**		
	# Classes	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Location-free	19	84.85	85.67	83.56	58.93	60.66	56.92
Kitchen	11	87.17	88.72	86.29	59.28	64.44	58.69
Living Room	4	91.43	91.91	90.61	71.66	75.28	71.04
Bathroom	4	93.75	93.75	92.98	53.24	57.47	52.55

patterns such as chopping, peeling, grating, or hair brushing. These activities were in some cases confusing to the model. Last but not least, television and background speech are very hard to distinguish from each other, especially when the television sound mainly contains speech. This was the case for P1 and P11 where P1 watched the news, and P11 watched a documentary. In order to further visualize the model performance, the overall confusion matrix across all LOPO evaluations is depicted in Figure 7.

7.1.2 Location-specific Modelling. Taking into account the fact that users can specify the location of the voice assistant during setup, we can build location-specific models by limiting classes to their respective context. Table 1 lists the activities for every location. Although there are several activities that can be considered as location-free classes (i.e., can happen anywhere), we followed the listing based on how the activities were split during the study. The majority of activities (11 activities) belong in the kitchen, whereas only 4 activities belong in the living room and bathroom. Table 3 lists the performance of each of the per-context models. Each model is

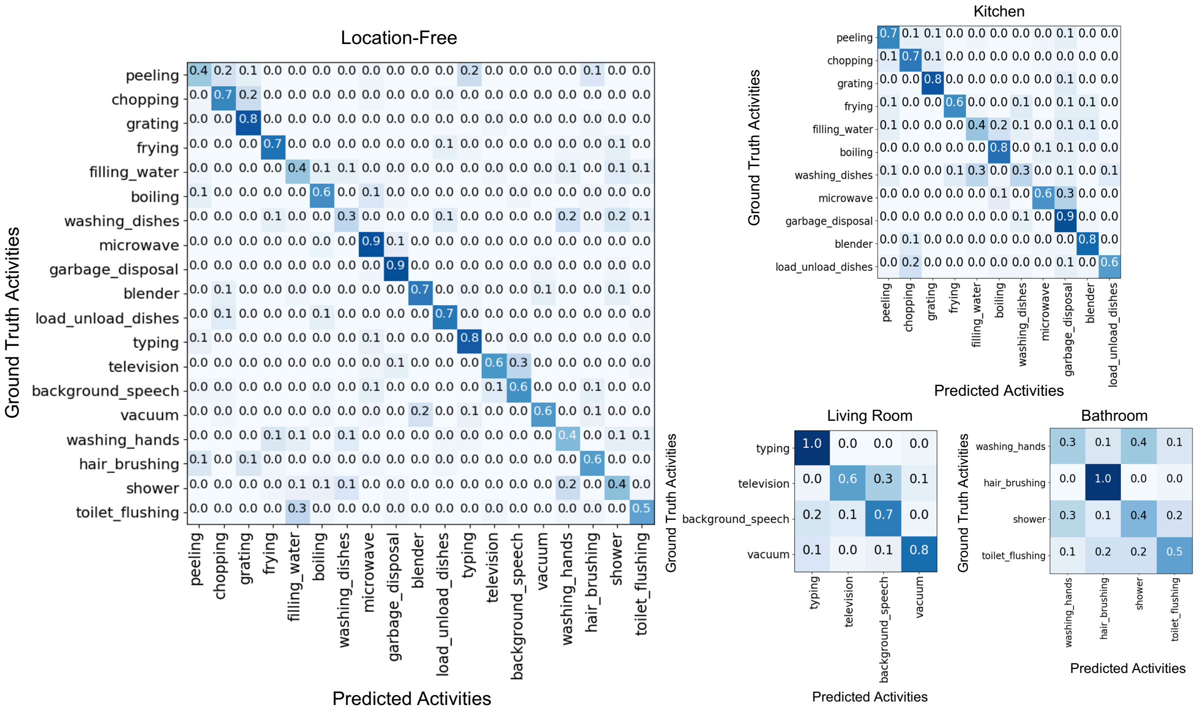


Fig. 8. Confusion matrices for activity recognition using mid-interaction segment as input

trained and tested on its corresponding set of activities. Within each context, typical mispredictions that were observed included television and background speech, water-related activities such as washing hands and shower, or washing dishes and filling water, etc. This can be observed in the corresponding confusion matrices in Figure 7. From location-specific modelling, we observe that reducing the number of classes clearly improves performance and allows the model to better distinguish classes, especially classes that might share some similar acoustic properties.

7.2 Mid-interaction Segment as Input

As described previously, the mid-interaction segment is the short audio gap between the user's query and the voice assistant's response.

7.2.1 Location-free Modelling. Similar to Section 7.1.1, we apply a LOPO evaluation using the mid-interaction segments of our data. The model was able to identify the set of 19 classes with an average precision of 58.93%, average recall of 60.66%, and an average F1-score of 56.92%. As expected, performance dropped compared to using the whole audio clip since the amount of acoustic information captured dropped as well. However, when put in perspective, the audio clips were reduced from 30 seconds to about 2-3 seconds of data. This is approximately an order of magnitude reduction which caused around 25% drop in precision/recall/F1-score. Moreover, looking at the overall confusion matrix in Figure 8, we observe that the model is still able to accurately classify several activities such as microwave, garbage disposal, typing, grating, etc. We observe there is more confusion in water-related activities compared to using whole audio clips.

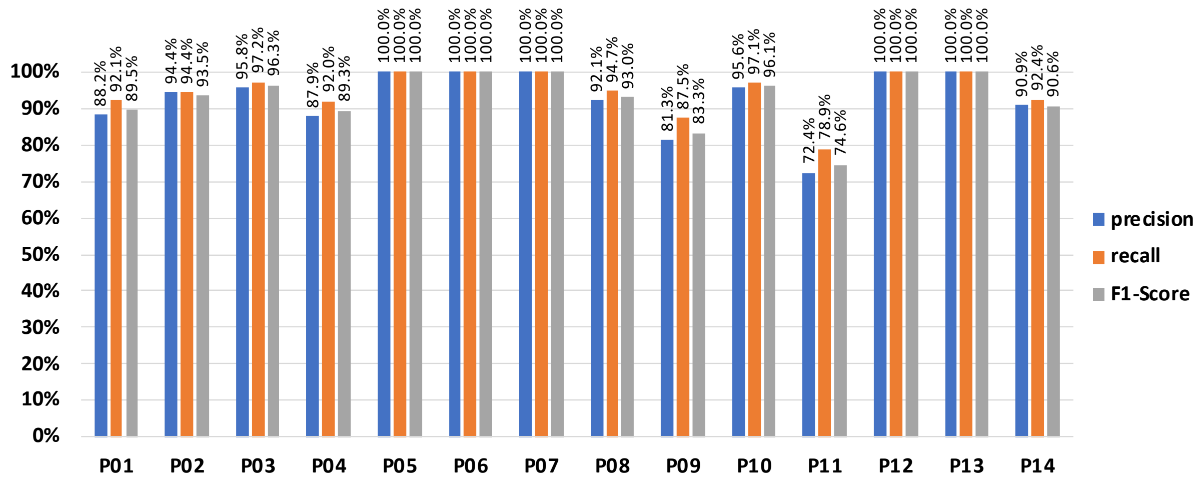


Fig. 9. Bar plot showing the performance scores of every personalized LOSO evaluation of the location-free model using whole audio clips

7.2.2 Location-specific Modelling. When moving to location-specific modelling, we observed a slight increase in performance for kitchen activities and a significant increase for living room activities (Table 3). However, for bathroom activities, performance slightly dropped. This can be explained by the greater overlap in water-related activities found in the bathroom. Looking at the corresponding confusion matrix in Figure 8, 75% of bathroom activities include water sounds that can be mispredicted. However, for living room activities, we observe a better distinction across activities, with some confusion between television and background speech.

7.3 Personalized Analysis

Every user's environment varies significantly whether in terms of background noise from the ambient hum of HVAC, or bustling traffic, or in terms of appliance sounds (microwave, blender, etc.) and device placement. For example, for several participants, the HVAC sound was very loud making it challenging to capture clear peeling or typing sounds. With some participants, the sound of a dog barking, a family member talking on the phone in another room, television playing in another room, baby crying, footsteps of a person coming down the stairs, participant coughing, door closing, phone notification ping, etc. were captured in the background. Except for the persistent HVAC sound, these background noises were not always present in all recordings. Applying personalized analysis can further improve performance and create models personalized to every user.

We investigated this personalization by investigating two approaches. The first approach consists of a Leave-One-Session-Out (LOSO) evaluation per user. In our data collection, participants performed every activity twice, which can be split into 2 separate sessions (each with 19 activities). Thus, for every user, we investigated building personalized models by training on one session and testing on the other. Considering training and testing per person, there is very limited amount of data, but it is important to keep in mind that personalization can be achieved over time. To that end, we ran such LOSO analysis using whole audio clips per user and demonstrated an average performance of 92.76% precision, 94.74% recall, and 93.29% F1-score. Figure 9 shows the model's performance per LOSO evaluation. Due to the limited training data per user in the LOSO setting, we investigated another approach that consisted of augmenting our training data per user with data from all other participants;

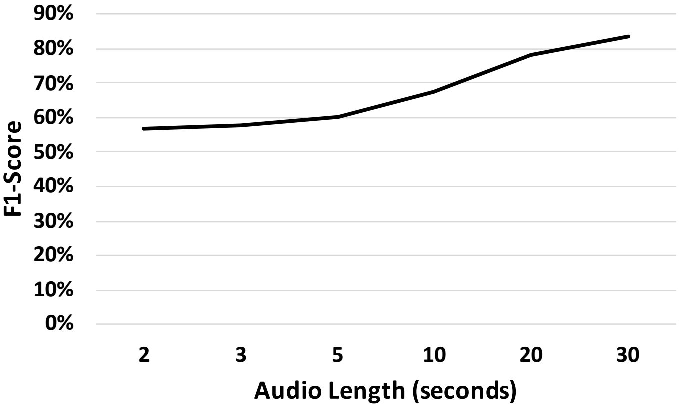


Fig. 10. F1-score performance with variable audio length

i.e. for every target user, the training data consisted of data from all other users in addition to data from one session of the target user, while the test data consisted of data from the other session. This approach effectively allowed us to increase our training data while still personalizing the model to every user. This can be considered similar to how voice assistants request repeating the trigger phrase a few times during setup to personalise the voice recognition model to a specific user. With this approach and using whole audio clips, we achieved an average performance of 91.94% precision, 94.27% recall, and 92.65% F1-score which is comparable to the LOSO performance.

8 DISCUSSION

8.1 Out-of-Scope Sounds

The evaluation procedures presented in this paper were done with a closed set of activities already observed by the classifier. However, in the real world, voice assistant commands are likely to be accompanied by "unknown" unclassified sounds or no sounds. Thus, evaluating the classifier using out-of-scope "unknown" classes can offer additional insights into how the system will handle such real-world situations.

To that end, we collected additional audio data, from 4 participants (not part of the study), of voice-based interactions accompanied by "unknown" background sounds, that mainly included no sounds as well as a few instances of dog barking, footstep sounds, and washing machine. The data collection process in this case was not scripted, and participants were given the devices to use freely. This resulted in 85 audio clips. Using the model trained on all the audio data collected in the study, we evaluated it on the "unknown" samples and observed that 78.8% of the instances were predicted as boiling. All instances had an average prediction confidence of 10.5% (± 18.5%). Boiling sounds collected during the study generally included a low bubbling sound which can be found similar to "no sounds". From our previous LOPO evaluation on our observed activities set, the prediction confidence value averaged around 83.9% (± 25.2%) which exceeds the average prediction confidence for the "unknown" instances. Thus, this can be used to define a confidence threshold for classifying and ignoring "unknown" sounds, i.e. a sound is classified as "unknown" if the top predicted class does not exceed a certain confidence threshold.

8.2 Recognition Performance vs. Audio Length

In section 7, we evaluated the ability of our framework to recognize activities from 30-second audio clips and 2-second audio clips (mid-interaction segment). To visualize how performance changes with varying audio length, we ran location-free LOPO analysis on audio clips with variable length. Recall that our audio clips contain around 7 seconds of audio before the interaction. We ran this analysis by looking at the audio segments starting from when the interaction occurs, which as whole is around 20 seconds. We decreased the audio length and observed how performance is affected (Figure 10). We observed a gradual increase in performance as the audio length increases; unsurprisingly, longer audio clips capture more acoustic properties that enable the model to better learn the acoustic patterns to distinguish activities. But it is worth noting that the improved performance observed for longer audio segments is also related to the annotation granularity of the data. The AudioSet used for training the feature extractor contained 10-second audio clips that are labeled at that granularity, despite acoustic events happening intermittently within the clip. This is natural due to the way acoustic events occur; e.g. chopping activity occurs intermittently. Thus, the CNN pretrained on AudioSet learned embeddings associated with weakly labeled 10-second segments. This is similar to our annotation process during the study, wherein we label the whole 30-second clip as one acoustic event despite the clip containing intermittent sounds as well as speech sounds. Thus, when taking shorter audio segments of the interaction such as 3-second clips, labeling them as the original label might not work as the segment might only contain speech. On the other hand, we posited that the drop in performance with shorter audio clips is also attributed to the fact that the audio signal is overlapped with speech. Specifically, the short audio clips tend to contain more speech than background sounds. Analysis of background sounds in the presence of speech is challenging to the inference model due to the mixed signal.

8.3 Voice Interaction Masking

To reiterate, the 30-second audio clip captures background sounds as well as the user's interaction with the voice assistant, which includes the user query and the voice assistant's answer. In our analysis so far, we did not deal with the human or machine speech present in the audio clips. Given that our current focus is on background sounds, we investigate the effect of removing the voice interaction from the audio clips on the model's ability to classify activities. To that end, we initially explored stripping the voice bands, that typically fall in the range of 80Hz - 3KHz, from the audio input [47]. This was done by applying a band-stop filter that eliminates the human voice. However, this approach proved challenging because, while the human speech content became unrecognizable, the machine speech was still recognizable and was just attenuated to a whisper. Moreover, the quality of the background sounds was reduced. In order to better eliminate both human and machine speech present in the foreground while preserving the original quality of background sounds, we explored a simple background-foreground separation method that is based on the REpeating Pattern Extraction Technique (REPET) method [37]. This method assumes the background has repeating elements and uses a similarity matrix and median filtering to compute a time-frequency masking of the spectrogram. Although this method was proposed for the task of music-voice separation, we investigated its use for our application. Initially, we applied this method to the whole 30-second audio clip. However, we observed that for some activities that include speech sounds in the background, e.g., mainly background conversation and television, this method masked background speech sounds as well. For our purposes, our focus was to apply a masking of the voice interaction only while preserving background sounds as much as possible. Therefore, we then reduced the time range to which we apply the masking method to the range where the voice interaction falls, which in most cases was on average between the 7^{th} second and the $15 - 20^{th}$ second of the whole 30 second clip. Figure 11 shows the log-mel spectrogram of a sample audio clip before and after masking the voice interaction. Using the background log-mel spectrograms, we repeated the same location-free LOPO evaluation and observed a performance of 78.63% average precision, 78.99% average recall, and 76.81% average F1-score, an approximate drop of 6% in performance compared to using

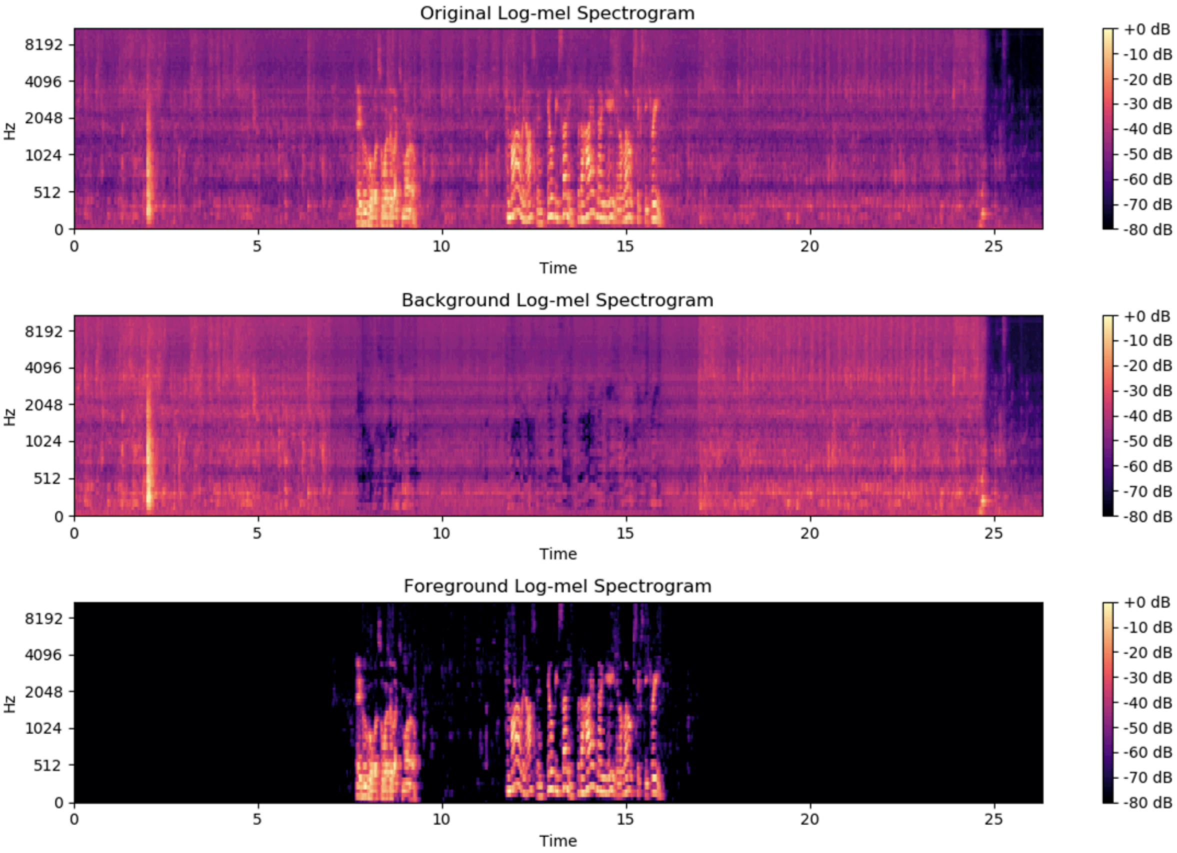
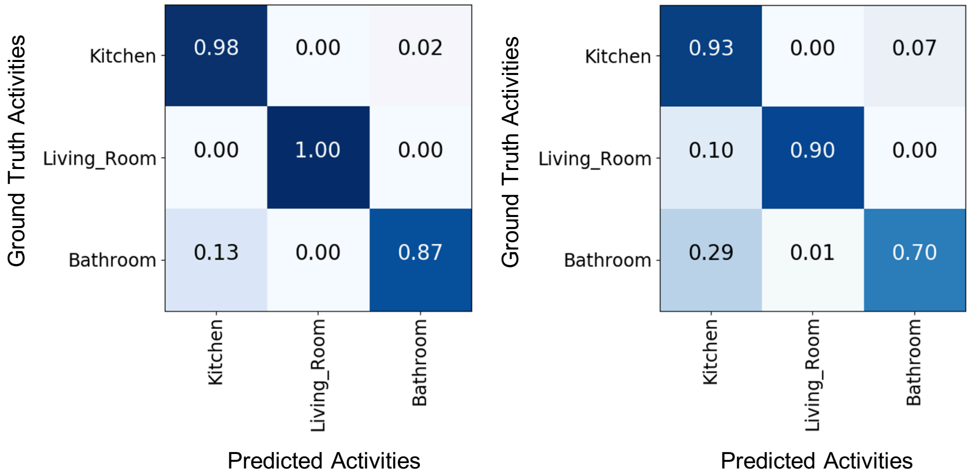


Fig. 11. Log-mel spectrograms of a sample audio clip before and after masking the voice interaction. Time-frequency masking applied to the original spectrograms results in background and foreground (voice interaction) separation. The background log-mel spectrograms are then used to train and evaluate our activity recognition framework.

the original audio without voice interaction masking. This drop in performance can be attributed to the fact that masking the voice interaction compromised the audio content leaving a blank spot in the spectrograms used as input to our recognition framework, as can be see in Figure 11.

8.4 Location Context Inference

With voice assistants, users have the ability to specify the location of the device during setup. This allows location-specific classification. However, it is often the case that users fail to specify the location or even change the device's location after some time. Thus, automatically inferring the physical context of the voice assistant would be beneficial for enabling location-specific classification without user intervention. Moreover, inferring the location of the device can help provide more context to the voice assistant thus ultimately improving their use and creating new experiences tailored to specific locations. In order to achieve that, activity classes predicted by our system can act as proxies for the location context. To illustrate, if the system predicts a series of activities typically performed in the kitchen (e.g. chopping, frying, etc.), we can infer that the device is placed in the kitchen.



(a) Whole 30s audio clip as input

(b) Mid-interaction segment as input

Fig. 12. Confusion Matrices for location context inference

In order to simulate this experiment, we applied a LOPO evaluation wherein we tested our location-free model, trained to classify all 19 classes, on 4 randomly selected clips from the hold-out participant for each context (kitchen, living room, and bathroom). The model predicts the activity in each clip individually, and then a majority voting scheme is applied to the set of predicted activities for each context in order to infer the device's location. For each participant, we repeated this process 10 times with randomly sampled clips within each context.

Using whole 30-second audio clips as input, on average, we achieved 96.81% precision, 95.24% recall, and 94.76% F1-score in location context inference. This shows the possibility to automatically infer the device's context after setup by utilizing the predicted activities captured during the first few interactions. Further investigating the results, we observed that kitchen and bathroom were sometimes confused with each other due to several classes in both locations belonging to water-related activities. Figure 12 depicts the overall confusion matrix of all the repeated iterations combined.

Additionally, the same experiment was implemented using the mid-interaction segment as input. As was already observed, when constrained with those short audio clips, activity recognition performance drops. This was also observed in location inference wherein, on average, a precision of 89.80%, a recall of 83.81%, and an F1-score 84.13% were achieved. Similarly, the confusion matrix is plotted in Figure 12.

8.5 Comparison to Prior Work

Several prior efforts have evaluated audio-based activity recognition systems. Due to differences in experimental design and evaluation metrics, it is difficult to compare them against each other in terms of performance. In this section, we present a more focused discussion of this prior work in light of the results we obtained in our studies.

As mentioned in Section 2.3, Laput *et al.* applied audio augmentation to the AudioSet data using sound effect libraries to build an audio context classification framework. Their best model, trained on this augmented data, achieved 80.4% overall accuracy for classifying 30 activity classes across 7 location contexts recorded in the wild. When trained using the AudioSet without augmentation, their overall accuracy dropped to 69.5% when tested on their real-world data. Liang *et al.* presented a similar work, in which an audio-based activity recognition framework was trained using 519,270 AudioSet embeddings extracted from a VGGish network pre-trained on the Youtube-100M dataset. They further showed that random oversampling of those embeddings prior to training the classifier boosts their overall accuracy. Thus, testing their model on a real-world data collected from 14 participants, they achieved an average top-1 accuracy of 64.16% when classifying 15 activity classes. They further applied a top-3 classification scheme which resulted in 83.6% average accuracy.

Similar to prior work, we leverage transfer learning by utilizing a model pre-trained on AudioSet as a feature extractor. However, unlike prior work, we train our classifier on our real-world data instead of an augmented or oversampled version of AudioSet. Thus, our training data does not exceed 500 samples (around 4 hours of data). Moreover, we are constrained by the question/answer mode of the voice assistant which restricted the length of our audio clips captured, and thus the amount of relevant acoustic information. Keeping in mind that the data is only captured when a user interacts with a voice assistant, every clip, therefore, includes this speech-based interaction which can make the recognition task more challenging. Despite the limited amount of data, we were able to achieve comparable and even better results compared to prior work when classifying 19 classes across 3 location contexts, achieving 85.7% average accuracy when training using whole audio clips in the location-free setting. This slight improvement in performance can also be attributed to the additional complexity of the pre-trained CNN model used compared to pre-trained VGGish network used in prior work. Our recognition framework includes two additional convolutional blocks and applies an entire audio clip for training as opposed to splitting the audio clip into segments.

8.6 Privacy Considerations

Audio has been extensively and successfully used as a sensing modality in activity recognition [24, 26]. However, while ambient audio has the potential to support contextually-relevant and finely-grained activity sensing, audio recording in the home runs the risk of capturing privacy-sensitive data. Naturally, the more data is captured, the more opportunities exist for identifying relevant signatures of behavior and context. On the other hand, extensive audio recording increases the chance that privacy issues emerge. We addressed this optimization problem by minimizing data collection as much as possible, and thereby prioritizing privacy. As previously discussed, the audio we obtained and analyzed was restricted to the sounds captured during the voice-based interactions with the assistant only. Nonetheless, even simply recording people's interactions with voice assistants can still be a concern, especially given people's perception of smart speakers in general. Prior work have shown that users have mixed opinions in regards to the device's ability to learn things about them, with some perceiving the benefits of such a feature depending on the context, which follow established theories presented in Contextual Integrity [1, 2, 30]. Additional steps can be taken in an effort to further mitigate privacy concerns. Applying speech filtering or voice masking techniques, as presented in Section 8.3, can help remove sensitive speech data. In terms of awareness and transparency mechanisms, it is possible to provide users with privacy notices, help them better understand risk-benefit trade-offs, and let them set privacy controls.

8.7 Limitations

Our evaluations show promising results in making use of voice-based interactions with conversational assistants to recognize activities of daily living even when restricted by the amount of audio data we are able to capture. That said, our work also has noteworthy limitations.

In our semi-naturalistic study, we collected data from only 14 participants, and they were explicitly instructed to perform a set of activities. However, our set of participants was highly diverse, representing individuals from all age groups and gender. Additionally, data collection occurred in participants' own homes, using their own tools, e.g., pots and pans. Moreover, participants were asked to perform the activities naturally, as if they were not part of a study. This allowed us to capture more variability and personalized data for each user.

Another limitation of the study was the lack of simultaneous activities. Real world settings can often be chaotic as people multitask, with multiple sounds occurring at the same time. Our system and evaluations focused on one activity at a time, and thus, did not include non-overlapping sounds originating from multiple activities. In a noisier environment, accuracy would suffer, as was observed with some participants' data. Taking advantage of sound augmentation methods for creating complex sound mixtures might be one way to improve the robustness of acoustic classification models.

9 FUTURE WORK

There are numerous opportunities for extending our work. As the pace of technical development in speech processing continues, conversational assistants are evolving to allow for more natural (voice) interactions, breaking away from the nearly exclusive question/answer mode we have today. This has already been seen in Google Assistants with the "Continued Conversation" feature that allows back-and-forth interactions [6]. With this feature, the mic listens for requests for up to 8 seconds, enabling users to have a more natural interaction with the voice assistant without having to say the wake word prior to each command. In future work, we plan to investigate how this continuous mode of interaction can help improve the performance of our approach, and support other types of recognition.

A highly promising research direction that we did not explore in this paper is the analysis of the *content* of exchanges between people and their assistants. While our work only focused on background sounds captured during interactions with the voice assistant, a person's activity could also be inferred from the type of query. If a user asks a question about a recipe, they are likely to be cooking. The semantic analysis of speech interaction is highly complementary to the acoustic-based methods presented in this work.

Lastly, we believe that voice-based interactions with conversational assistants can enable many context-driven applications in various domains, and particularly in healthcare. For example, by extracting acoustic biomarkers from audio interactions, assistants could identify speech and activity patterns associated with cognitive impairments such as dementia. The devices could play even more active roles, such as by acting as virtual caregivers, reminding seniors of their daily tasks, medication schedule, and serving as the foundation for home-based interventions.

10 CONCLUSION

Voice-based interactions with conversational assistants capture background environmental sounds that can be a rich source of context. In this paper, we investigate the potential use of such data to recognize essential activities of daily living. We conducted a semi-naturalistic study in real-homes using an *add-on* accessory for the Google Home device. Critically, the accessory collected audio data without affecting the assistant's main functionalities. In the first step of our approach, the data was processed using a recognition framework that utilizes a pre-trained model on AudioSet as a feature extractor. Secondly, we trained a classifier to recognize 19 common home-related activities. We evaluated our system in both location-free and location-specific settings with promising performance and robustness to environmental variability. We achieved 83.56% F1-score in location-free evaluations with an increase in performance for location-specific evaluations going up to 92.98% for bathroom activities. In sum, this research represents a step forward in building systems that recognize activities of daily living in real-world settings while leveraging the ubiquity of conversational assistants in the home.

ACKNOWLEDGMENTS

This work was funded in part by a Google Faculty Research Award.

REFERENCES

[1] 2020. Smart Home Personal Assistants: A Security and Privacy Review. *Comput. Surveys* (22 July 2020).

[2] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. 2019. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/soups2019/presentation/abdi

[3] M'hamed Bilal Abidine and Belkacem Fergani. 2015. News Schemes for Activity Recognition Systems Using PCA-WSVM, ICA-WSVM, and LDA-WSVM. *Information* 6, 3 (2015), 505–521. https://doi.org/10.3390/info6030505

[4] A. Arcelus, R. Goubran, H. Sveistrup, M. Bilodeau, and F. Knoefel. 2010. Context-aware smart home monitoring through pressure measurement sequences. In *2010 IEEE International Workshop on Medical Measurements and Applications*. 32–37.

[5] 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 (Sept. 2018), 24 pages. https://doi.org/10.1145/3264901

[6] John Callaham. 2018. Speaking to Google Home will now be more natural with Continued Conversation. https://www.androidauthority.com/google-home-continued-conversation-878770/

[7] Hongzhao Chen, Guijin Wang, Jing-Hao Xue, and Li He. 2016. A novel hierarchical framework for human action recognition. *Pattern Recognition* 55 (2016), 148 – 159. https://doi.org/10.1016/j.patcog.2016.01.020

[8] K. S. Chun, H. Jeong, R. Adaimi, and E. Thomaz. 2020. Eating Episode Detection with Jawbone-Mounted Inertial Sensing. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 4361–4364. https://doi.org/10.1109/EMBC44109.2020.9175949

[9] Keum San Chun, Ashley B. Sanders, Rebecca Adaimi, Necole Streeper, David E. Conroy, and Edison Thomaz. 2019. Towards a Generalizable Method for Detecting Fluid Intake with Wrist-Mounted Sensors and Adaptive Segmentation. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 80–85. https://doi.org/10.1145/3301275.3302315

[10] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer. 2016. Monitoring Activities of Daily Living in Smart Homes: Understanding human behavior. *IEEE Signal Processing Magazine* 33, 2 (March 2016), 81–94. https://doi.org/10.1109/MSP.2015.2503881

[11] Svilen Dimitrov, Jochen Britz, Boris Brandherm, and Jochen Frey. 2014. Analyzing Sounds of Home Environment for Device Recognition. In *Ambient Intelligence*, Emile Aarts, Boris de Ruyter, Panos Markopoulos, Evert van Loenen, Reiner Wichert, Ben Schouten, Jacques Terken, Rob Van Kranenburg, Elke Den Ouden, and Gregory O'Hare (Eds.). Springer International Publishing, Cham, 1–16.

[12] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is It OK If I Eat You?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[13] Daniel J. Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi. 01 Oct. 2020. When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers. *Proceedings on Privacy Enhancing Technologies* 2020, 4 (01 Oct. 2020), 255 – 276. https://doi.org/10.2478/popets-2020-0072

[14] M. Forouzanfar, M. Mabrouk, S. Rajan, M. Bolic, H. R. Dajani, and V. Z. Groza. 2017. Event Recognition for Contactless Activity Monitoring Using Phase-Modulated Continuous Wave Radar. *IEEE Transactions on Biomedical Engineering* 64, 2 (2017), 479–491.

[15] Radhika Garg and Subhasree Sengupta. 2020. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 11 (March 2020), 24 pages. https://doi.org/10.1145/3381002

[16] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA.

[17] Google. [n.d.]. Introduction to the Google Assistant Service | Google Assistant SDK. https://developers.google.com/assistant/sdk/guides/service/python

[18] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135.

[19] Peter Hevesi, Sebastian Wille, Gerald Pirkl, Norbert Wehn, and Paul Lukowicz. 2014. Monitoring Household Activities and User Location with a Cheap, Unobtrusive Thermal Sensor Array. 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, 141–145. https://doi.org/10.1145/2632048.2636084

[20] Bret Kinsella. 2019. Loup Ventures Says 75% of U.S. Households Will Have Smart Speakers by 2025, Google to Surpass Amazon in Market Share. https://voicebot.ai/2019/06/18/loup-ventures-says-75-of-u-s-households-will-have-smart-speakers-by-2025-google-to-surpass-amazon-in-market-share/

[21] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA) *(CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/2854946.2854961

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

[23] Dounia Lahoual and Myriam Frejus. 2019. When Users Assist the Voice Assistants: From Supervision to Failure Resolution. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, Article CS08, 8 pages. https://doi.org/10.1145/3290607.3299053

[24] 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

[25] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos. 2016. VLAD3: Encoding Dynamics of Deep Features for Action Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1951–1960.

[26] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 17 (March 2019), 18 pages. https://doi.org/10.1145/3314404

[27] Martino Lombardi, Roberto Vezzani, and Rita Cucchiara. 2015. Detection of Human Movements with Pressure Floor Sensors. In *ICIAP*.

[28] Raspberry Pi Camera Module. [n.d.]. Raspberry Pi Camera Module. https://www.raspberrypi.org/documentation/usage/camera/

[29] S. C. Mukhopadhyay. 2015. Wearable Sensors for Human Activity Monitoring: A Review. *IEEE Sensors Journal* 15, 3 (March 2015), 1321–1330. https://doi.org/10.1109/JSEN.2014.2370945

[30] Helen Nissenbaum. 2004. Privacy As Contextual Integrity. *Washington Law Review* 79 (05 2004).

[31] NPR. 2020. NPR and Edison Research Report: 60M U.S. Adults 18 Own a Smart Speaker. https://www.npr.org/about-npr/794588984/npr-and-edison-research-report-60m-u-s-adults-18-own-a-smart-speaker

[32] Shigeyuki Odashima, Toshikazu Kanaoka, Katsushi Miura, Keiju Okabayashi, and Naoyuki Sawasaki. 2016. Human Activeness Recognition by Variety of Rare Sounds. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 181–184. https://doi.org/10.1145/2968219.2971400

[33] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[34] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[35] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information": Personification and Ontological Categorization of Smart Speaker-Based Voice Assistants by Older Adults. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 214 (Nov. 2019), 21 pages. https://doi.org/10.1145/3359316

[36] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 459, 13 pages. https://doi.org/10.1145/3173574.3174033

[37] Zafar Rafii and Bryan Pardo. 2012. Music/Voice Separation Using the Similarity Matrix. In *ISMIR*. 583–588.

[38] Simon Robinson, Jennifer Pearson, Shashank Ahire, Rini Ahirwar, Bhakti Bhikne, Nimish Maravi, and Matt Jones. 2018. Revisiting "Hole in the Wall" Computing: Private Smart Speakers and Public Slum Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 498, 11 pages. https://doi.org/10.1145/3173574.3174072

[39] Bradley Spicer. 2020. How Much Internet Speed Does Your Smart Home Need? https://www.smarthomebit.com/how-much-internet-speed-does-your-smart-home-need/

[40] Seeed Studio. [n.d.]. ReSpeaker 4-Mic Linear Array Kit for Raspberry Pi. http://wiki.seeedstudio.com/ReSpeaker_4-Mic_Linear_Array_Kit_for_Raspberry_Pi/

[41] Madiha Tabassum, Tomasz Kosiundefinedski, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 153 (Dec. 2019), 23 pages. https://doi.org/10.1145/3369807

[42] Catherine Tong, Shyam A. Tailor, and Nicholas D. Lane. 2020. Are Accelerometers for Activity Recognition a Dead-end? arXiv:2001.08111 [cs.CV]

[43] Sébastien Tremblay, Dany Fortin-Simard, Erika Blackburn-Verreault, Sébastien Gaboury, Bruno Bouchard, and Abdenour Bouzouane. 2015. Exploiting Environmental Sounds for Activity Recognition in Smart Homes. In *AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*.

[44] Toshifumi Tsukiyama. 2015. In-home Health Monitoring System for Solitary Elderly. *Procedia Computer Science* 63 (2015), 229 – 235. https://doi.org/10.1016/j.procs.2015.08.338 The 6th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2015)/ The 5th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2015)/ Affiliated Workshops.

[45] Prabitha Urwyler, Luca Rampa, Reto Stucki, Marcel Büchler, René Martin Müri, Urs Peter Mosimann, and Tobias Nef. 2015. Recognition of activities of daily living in healthy subjects using two ad-hoc classifiers. In *Biomedical engineering online*.

[46] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, and S. Méniard. 2011. The sweet-home project: Audio technology in smart homes to improve well-being and reliance. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 5291–5294.

[47] Wei Wang, Fatjon Seraj, Nirvana Meratnia, and Paul J. M. Havinga. 2019. Privacy-Aware Environmental Sound Classification for Indoor Human Activity Recognition. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '19)*. Association for Computing Machinery, New York, NY, USA, 36–44. https://doi.org/10.1145/3316782.3321521

[48] Jiaxuan Wu, Yunfei Feng, and Peng Sun. 2018. Sensor Fusion for Recognition of Activities of Daily Living. In *Sensors*.

[49] Che-Chang Yang and Yeh-Liang Hsu. 2012. Remote monitoring and assessment of daily activities in the home environment. *Journal of Clinical Gerontology and Geriatrics* 3, 3 (2012), 97 – 104. https://doi.org/10.1016/j.jcgg.2012.06.002