SoundWatch: Deep Learning for Sound Accessibility on Smartwatches

By Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Khoa Nguyen, Rachel Grossman-Kahn, Leah Findlater, and Jon Froehlich

Abstract

Smartwatches have the potential to provide glanceable, always-available sound feedback to people who are deaf or hard of hearing (DHH). We present SoundWatch, a smartwatch-based deep learning application to sense, classify, and provide feedback about sounds occurring in the environment. To design SoundWatch, we first examined four low-resource sound classification models across four device architectures: watch-only, watch+phone, watch+phone+cloud, and watch+cloud. We found that the best model, VGG-lite, performed similar to the state of the art for nonportable devices although requiring substantially less memory ($\sim 1/3^{rd}$) and that the watch+phone architecture provided the best balance among CPU, memory, network usage, and latency. Based on these results, we built and conducted a lab evaluation of our smartwatch app with eight DHH participants. We found support for our sound classification app but also uncovered concerns with misclassifications, latency, and privacy.

1. INTRODUCTION

Smartwatches have the potential to provide glanceable and always-available sound feedback to people who are deaf or hard of hearing (DHH) across multiple contexts.[3,5,17] A recent survey with 201 DHH participants[3] showed that, compared to smartphones and head-mounted displays (HMDs), a smartwatch is the most preferred device for nonspeech sound awareness due to privacy, social acceptability, and integrated support for both visual and haptic feedback.

Most prior work in wearable sound awareness, however, has focused on smartphones,[1,20] HMDs,[6,9] or custom wearable devices[13] that provide limited information (e.g., loudness) through a single modality (e.g., vision). For smartwatches specifically, studies have examined formative design prototypes for sound feedback,[5,17] but these prototypes have not included automatic sound classification—our focus. Furthermore, although recent deep learning research (e.g., see Jain et al.[11]) has examined models to automatically classify sounds, these cloud- or laptop-based models have high memory and processing power requirements and are unsuitable for low-resource portable devices.

Building on the above research, we present two smartwatch-based studies and a custom smartwatch-based application, called SoundWatch (see Figure 1). To design SoundWatch, we first quantitatively examined four state-of-the-art low-resource deep learning models for sound classification: *MobileNet, Inception, ResNet-lite,* and a quantized version of model used in *HomeSound,*[11] which we call VGG-lite, across four device architectures: watch-only, watch+phone, watch+phone+cloud, and watch+cloud. These approaches were intentionally selected to examine trade-offs in computational and network requirements, power efficiency, data privacy, and latency. Although direct comparison to prior work is challenging, our experiments show that the best classification model (VGG-lite) performed similar to the state of the art for nonportable devices although requiring substantially less memory ($\sim 1/3^{rd}$). We also observed a strict accuracy-latency trade-off: the most accurate model was the slowest. Finally, we found that the two phone-based architectures (watch+phone and watch+phone+cloud) outperformed the watch-centric designs (watch-only and watch+cloud) in terms of CPU, memory, battery usage, and end-to-end latency.

Based on these quantitative experiments, we built SoundWatch and conducted a qualitative lab evaluation with eight DHH participants. SoundWatch incorporates the best-performing classification model from our system experiments (VGG-lite) and, for the purposes of evaluation, can be switched between all four device architectures. During the 90-min study session, participants used our prototype in three locations on a university campus (a home-like lounge, an office, and outdoors) and took part in a semistructured interview about their experiences, their views on accuracy-latency trade-offs and privacy, and ideas and concerns for future wearable sound awareness technology. We found that all participants generally appreciated SoundWatch across all contexts, reaffirming past sound awareness work.[3,5] However, misclassifications were concerning, especially outdoors because of background noise. For accuracy-latency trade-offs, participants wanted minimum delay for urgent sounds (e.g., car honk and fire alarms)—to take any required action—but maximum accuracy for nonurgent sounds (e.g., speech and background noise) to not be unnecessarily disturbed. Finally, participants selected watch+phone as the most preferred architecture due to privacy concerns with the cloud, versatility (no Internet connection required), and speed (watch+phone was faster than watch-only).

In summary, our work contributes (1) a comparison of deep learning models for sound classification on mobile devices; (2) a new smartwatch-based sound identification

Figure 1. Different use cases of our *SoundWatch* sound classification app and one of the four architectures (watch+phone).

system, called SoundWatch, with support for four device architectures; and (3) qualitative insights from in situ evaluation with eight DHH users, such as reactions to our designs, architectures, and ideas for future implementations.

This paper is based on our earlier ASSETS paper.[12] Since that paper was accepted in June 2020, much has changed. We released the SoundWatch codebase as open source (https://bit.ly/3bvgCLI) and our work helped guide subsequent literature (e.g., see Guo et al.[7]). The SoundWatch app is now available publicly on the Google Play Store (https://bit.ly/3bpEPTF, 500+ downloads to date). Additionally, sound recognition is integrated into both the major mobile platforms: Apple iOS and Google Android, demonstrating the impact of our work.

2. RELATED WORK
We situate our work within sound awareness needs, sound awareness tools, and sound classification research.

2.1. Sound awareness needs
Formative studies have examined the sounds, sound characteristics, and feedback modalities desired by DHH users. For sounds of interest, two large-scale surveys[1,3] showed DHH people most prefer urgent and safety-related sounds (e.g., sirens) followed by appliance alerts (e.g., microwave beep) and sounds about the presence of people (e.g., door knock calls). These preferences may be modulated by cultural factors: people who prefer oral communication may be more interested in some sounds (e.g., phone ring and conversations) than those who prefer sign language.[1,3]

In addition to these sounds, DHH users tend to desire information about certain sound characteristics (e.g., identity, location, and time of occurrence) more than others (e.g., loudness, duration, and pitch).[5,15] However, the utility of these characteristics may vary by location. For example, at home, awareness of a sound's identity and location may be sufficient,[10,11] but directional indicators are more important when mobile.[5] Besides location, social context (e.g., friends vs. strangers) could influence the use of the sound awareness tool,[3] and thus offering options for customization is key (e.g., using a sound-filtering menu).

In terms of feedback modalities, studies suggest combining visual and vibrational information for sound awareness[5, 17]; a smartwatch can provide both. Within the two modalities, prior work recommends using vibration to notify about sound occurrence and vision to show more information[1, 10]—which we also explore—although a recent study showed value in using complex vibration patterns to convey richer feedback (e.g., direction).[5]

We build on the above studies by examining the use of working smartwatch prototypes across contexts and revealing qualitative reactions and suggestions for system design.

2.2. Sound awareness technologies
Early research in sound awareness studied wrist-worn vibrotactile solutions, primarily to aid speech therapy by conveying voice tone[22] or frequency[21]; this work is complementary to our focus on nonspeech sound awareness. More recent work has examined stationary solutions for nonspeech sound awareness, such as on desktop displays.[15] Though useful for specific applications, these solutions are not conducive to multiple contexts. Toward portable solutions, Bragg et al.[1] and Sicong et al.[20] used smartphones to recognize and display sound identity (e.g., phone ringing and sirens). However, they evaluated their app in a single context (office[1] or a deaf school[20]) and focused on user interface rather than system performance—both are critical to user experience.

Besides smartphones, wearable solutions such as HMDs[6,9] and wrist-worn devices[13] have been examined. For example, Gorman[6] and Kaneko et al.[13] displayed the location of sound sources on an HMD and a custom wrist-worn device, respectively. We explore smartwatches to provide sound identity, the most desired sound property by DHH users.[1, 15] Although not specifically focused on smartwatches, Jain et al.[11] examined smartwatches as complementary alerting devices to smarthome displays that sensed and processed sound information locally and broadcasted it to the watches; we examine a self-contained smartwatch solution.

In summary, although prior work has explored sound awareness tools for DHH people, such as on portable devices,[6,9,13] this work has not yet built and evaluated a working smartwatch-based solution—a gap we address in our work.

2.3. Sound classification research

Early efforts in classifying sounds relied on handcrafted features such as zero-crossing rate, frame power, and pitch.[14,18] Though they performed reasonably well on clean sound files, these features fail to account for acoustic variations in the field (e.g., background noise).[14] More recently, machine learning-based classification has shown promise for specific field tasks such as gunshot detection or intruder alert systems.[4] For broad use cases, deep learning-based solutions have been investigated.[11,20] For example, Sicong et al.[20] explored a lightweight convolutional neural network (CNN) on smartphones to classify nine sounds preferred by DHH users (e.g., fire alarm and doorbell) in a school setting. Jain et al.[11] used deep CNNs running on a tablet to classify sounds in the homes of DHH users, achieving an overall accuracy of 85.9%. We closely follow the latter approach in our work by adapting it to low-resource devices (phone and watch) and performing evaluations in multiple contexts (home, work, and outdoors).

3. THE SOUNDWATCH SYSTEM

SoundWatch is an Android-based app designed for commercially available smartwatches to provide glanceable, always-available, and private sound feedback in multiple contexts. Building on previous work,[5,11] SoundWatch informs users about three key sound properties: identity, loudness, and time of occurrence through customizable visual and vibrational sound alerts (see Figures 1 and 3). We use a deep learning-based sound classification engine (running on the watch, paired phone, or cloud) to continually sense and process sound events in real time. Here, we describe our sound classification engine, our privacy-preserving sound sensing pipeline, system architectures, and implementation. Our codebase is open sourced: https://bit.ly/3bvgCLI.

3.1. Sound classification engine

To create a robust, real-time sound classification engine, we followed an approach similar to *HomeSound*,[11] which uses transfer learning to adapt a deep CNN–based image classification model (VGG) for sound classification. We downloaded three recently released (in Jan 2020) image-classification networks for small devices: *MobileNet,* 3.4MB; *Inception,* 41MB; and *ResNet-lite,* 178.3MB, and we used the quantized version of the network in HomeSound,[11] which we call VGG-lite, 281.8MB. We hypothesized that each network would offer different accuracy and latency trade-offs.

To perform transfer learning, similar to Jain et al.,[11] we used a large corpus of sound effect libraries—each of which provides a collection of high-quality, prelabeled sounds. Samples for 20 common sounds preferred by DHH people (e.g., dog bark, door knock, and speech)[1,3] were downloaded from six libraries—BBC, Freesound, Network Sound, UPC, TUT, and TAU. All sound clips were converted to a single format (16KHz, 16-bit, mono) and silences greater than one second were removed, resulting in 35.6 h of recordings. We divided the sound classes into three categories (see Table 1): high priority (containing the three most desired sounds by DHH people[1,15]); medium-priority sounds (10 sounds); and all sounds (20 sounds). Finally, we used the method by Hershey et al.[8] to compute log mel-spectrogram features for each category, which were then fed to the four networks, generating three models for each architecture (12 in total).

3.2. Sound sensing pipeline

For always-listening apps, privacy is a key concern. Although SoundWatch relies on a live microphone, we designed our sensing pipeline to protect user privacy. The system processes the sound locally on the watch or phone and, in the case of the cloud-based architectures, only uploads low-dimensional mel-spectrogram features. Although these features can be used to identify speech activity, the spoken content is challenging to recover. For signal processing, we take a sliding window approach: the watch samples the microphone at 16KHz and segments data into 1-second

Figure 2. A diagram of the four SoundWatch architectures with their sensing pipelines. Block widths are for illustration only and do not indicate actual computation time.
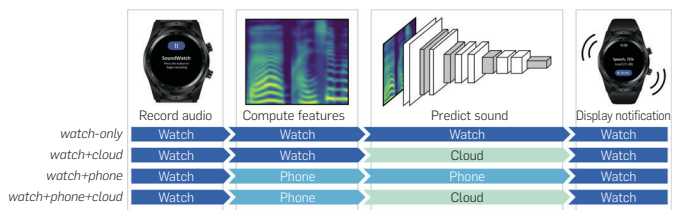

Figure 3. The SoundWatch user interface showing the (a) opening screen with a button to begin recording audio, (b) the notification screen with a "10-min" mute button, (c) the main app screen with more mute options, and (d) the paired phone app for customizing the list of enabled sounds.
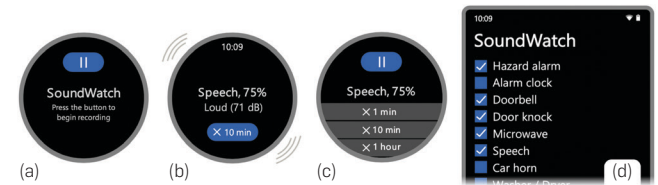
Table 1. The sounds and categories used to train our sound classification models.

All sounds (N = 20)	Fire/smoke alarm, alarm clock, door knock, doorbell, door-in-use, microwave, washer/dryer, phone ringing, speech, laughing, dog bark, cat meow, baby crying, vehicle running, car horn, siren, bird chirp, water running, hammering, drilling
High priority (N = 3)	Fire/smoke alarm, alarm clock, door knock
Medium priority (N = 10)	Fire/smoke alarm, alarm clock, door knock, doorbell, microwave, washer/dryer, phone ringing, car horn, siren, water running
Home context (N = 11)	Fire/smoke alarm, alarm clock, door knock, doorbell, door-in-use, microwave, washer/dryer, speech, dog bark, cat meow, baby crying
Office context (N = 6)	Fire/smoke alarm, door knock, door-in-use, phone ringing, speech, laughing
Outdoor context (N = 9)	Dog bark, cat meow, vehicle running, car horn, siren, bird chirp, water running, hammering, drilling

buffers (16,000 samples), which are fed to the sound classification engine. To extract loudness, we compute the average amplitude in the window. All sounds at or above 50% confidence and 45dB loudness are notified; others are ignored.

3.3. System architectures

We implemented four device architectures for SoundWatch: watch-only, watch+phone, watch+cloud, and watch+phone+cloud (see Figure 2). Because the sound classification engine (computing features and predicting sound) is resource intensive, the latter three architectures use a more powerful device (phone or cloud) for running the model. For only the cloud-based architectures, sound features are computed before being sent to the cloud to protect user privacy—that is, on the watch (watch+cloud) or on the phone (watch+phone+cloud). For communication, we use Bluetooth Low Energy (BLE) for watch-phone and WiFi or a cellular network for watch-cloud or phone-cloud.

3.4. User interface

For glanceability, we designed the SoundWatch app as a push notification; when a classified sound event occurs, the watch displays a notification along with a vibration alert. The display includes sound identity, classification confidence, loudness, and time of occurrence (see Figure 3). Importantly, each user can mute an alerted sound by clicking on the "10 min" mute button, or by clicking on the "open" button and selecting from a scroll list of mute options (1 min, 5 min, 10 min, 1 h, 1 day, or forever). Additionally, the user can filter alerts for any sounds using a customization menu on the paired phone app (see Figure 3d). Although future versions should run as an always-available service in Android, currently, the app must be explicitly opened on the watch (see Figure 3a). Once opened, the app runs continuously in the background.

4. SYSTEM EVALUATION

To assess the performance of our SoundWatch system, we performed two sets of evaluations: (1) a comparison of the four state-of-the-art sound classification models for small devices and (2) a comparison of the four architectures: watch-only, watch+phone, watch+cloud, and watch+phone+cloud. For all experiments, we used the Ticwatch Pro Android watch (4×1.2GHz, 1GB RAM) and the Honor 7x Android phone (8×2GHz, 3GB RAM). To emulate the cloud, we used an Intel i7 desktop running Windows 10 (4×2.5GHz, 16GB RAM).

4.1. Model comparison

We present our evaluation of classification accuracy and latency of the four models.

Accuracy. To calculate the "in-the-wild" accuracy of the models, we collected our own "naturalistic" dataset similar to *Home-Sound*.[11] We recorded 20 sound classes from nine locations (three homes, three offices, and three outdoors) using the same hardware as SoundWatch: Ticwatch Pro with a built-in microphone. For each sound class, we recorded three 10-second samples at three distances (5, 10, and 15 feet). When possible, we produced sounds naturally (e.g., by knocking or using a microwave). For certain difficult-to-produce sounds—such as a fire alarm—we played snippets of predefined videos on a laptop or phone with external

speakers (total 54 videos were used). In total, we collected 540 recordings (~1.5 h).
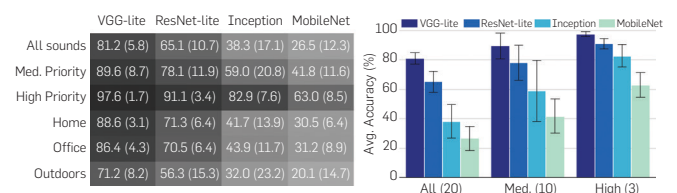
Before testing our model, we divided our recordings into the three categories (all sounds, high priority, and medium priority) similar to our training set (see Table 1). For the medium- and high-priority test-sets, 20% of the sound data was from excluded categories that our models should ignore (called the "unknown" class). For example, the high-priority test-set included 20% recordings from outside of the three high-priority classes (fire/smoke alarm, alarm clock, and door knock).

Figure 4 shows the results of classifying sounds in each category. Overall, VGG-lite performed best (avg. inference accuracy = 81.2%, SD = 5.8%) followed by ResNet-lite (65.1%, SD = 10.7%), Inception (38.3%, SD = 17.1%), and MobileNet (26.5%, SD = 12.3%); a one-way repeated measures ANOVA on all sounds yielded a significant effect of models on the accuracy ($F_{3,2156}$ = 683.9, $p < .001$). As expected, the inference accuracy increased as the number of sounds decreased from all (20 sounds) to medium (10 sounds) and high priority (3 sounds). In analyzing performance as a function of context, home and office outperformed outdoors for all models. With VGG-lite, for example, average accuracy was 88.6% (SD = 3.1%) for home, 86.4% (SD = 4.3%) for office, and 71.2% (SD = 8.2%) for outdoors. A post hoc inspection revealed that outdoor recordings incurred interference due to the background noise.

To assess interclass errors, we computed a confusion matrix for medium-priority sounds. Although per-class accuracy varied across models, microwave, door knock, and washer/dryer were consistently the best-performing classes, with VGG-lite achieving average accuracy of 100% (SD = 0), 100% (SD = 0), and 96.3% (SD = 2.3%), respectively. The worst-performing classes were more model dependent but generally included alarm clock, phone ring, and siren, with VGG-lite achieving 77.8% (SD = 8.2%), 81.5% (SD = 4.4%), and 88.9% (SD = 3.8%), respectively. For these poorly performing classes, understandable mix-ups occurred such as confusions among similar sounding events (e.g., alarm clocks and phone rings).

Latency. Low latency is crucial to achieving a real-time sound identification system. To evaluate model latency, we wrote a script to loop through the sound recordings in our dataset for 3 h (1080 sounds) and measured the time required to classify sounds from the input features on both the watch and the phone. Understandably, the latency increased with the model size: the smallest model, MobileNet, performed the fastest on both devices (*avg. latency* on watch: 256 ms, SD = 17 ms; phone: 52 ms, SD = 8 ms), followed by Inception (watch: 466 ms, SD = 15 ms; phone: 94 ms, SD = 4 ms), and

Figure 4. Average accuracy (and SD) of the four models for three sound categories and three contexts. Error bars in the graph show 95% confidence intervals.

	VGG-lite	ResNet-lite	Inception	MobileNet
All sounds	81.2 (5.8)	65.1 (10.7)	38.3 (17.1)	26.5 (12.3)
Med. Priority	89.6 (8.7)	78.1 (11.9)	59.0 (20.8)	41.8 (11.6)
High Priority	97.6 (1.7)	91.1 (3.4)	82.9 (7.6)	63.0 (8.5)
Home	88.6 (3.1)	71.3 (6.4)	41.7 (13.9)	30.5 (6.4)
Office	86.4 (4.3)	70.5 (6.4)	43.9 (11.7)	31.2 (8.9)
Outdoors	71.2 (8.2)	56.3 (15.3)	32.0 (23.2)	20.1 (14.7)

ResNet-lite (watch: 1615 ms, *SD* = 30 ms; phone: 292 ms, *SD* = 13 ms). VGG-lite, the largest model, was the slowest (watch: 3397 ms, *SD* = 42 ms; phone: 610 ms, *SD* = 15 ms).

Model comparison summary. In summary, for phone and watch models, we observed a strict accuracy–latency trade-off—for example, the most accurate model VGG-lite (*avg. accuracy* = 81.2%, *SD* = 5.8%) was also the slowest (*avg. latency* on watch: 3397 ms, *SD* = 42 ms). Further, the models MobileNet and Inception performed too poorly for practical use (*avg. accuracy* < 40%). ResNet-lite was in the middle (*avg. accuracy* = 65.1%, *SD* = 10.7%; *avg. latency* on watch: 1615 ms, *SD* = 30 ms).

Comparison to prior approach. We also evaluated the performance of the full VGG model running on the cloud, which is used in the state-of-the-art prior work on sound classification.[11] The average inference accuracy (84.4%, *SD* = 5.5%) was only slightly better than our best mobile-optimized model (VGG-lite, *avg.*= 81.2%, *SD* = 5.8%)—a promising result as our VGG-lite model is less than 1/3rd the size of VGG (281.8MB vs. 845.5MB).

4.2. Architecture evaluation

We compared the performance of four different architectures of SoundWatch: watch-only, watch+phone, watch+cloud, and watch+phone+cloud (see Figure 2), which may differ in performance and usability.

For each architecture, we used the most accurate model on the watch and phone: VGG-lite; the cloud used the full VGG model. Informed by prior work,[16] we measured CPU, memory, network usage, end-to-end latency, and battery consumption. For the evaluation, we used a script running on a laptop that looped through the sound recordings for 3 h to generate sufficient sound samples (1080). For the battery experiment only, the script ran until the watch battery reached 30% or less (i.e., just above the 25% trigger for low-power mode), a common evaluation approach.[16] To determine CPU, memory, and network usage, we used *Android Profiler,* a commonly used profiling tool. For power usage, we used *Battery Historian*. Finally, to determine end-to-end latency, we measured the elapsed time (in milliseconds) between the start of the sound recording to when the notification is shown. Here, we detail our results.

CPU Utilization. Minimizing CPU use will maximize the smartwatch's battery performance and lower the impact on other running apps. Our results for CPU usage on the watch and phone are as shown in Figure 5a. As expected, the watch's CPU utilization was lowest for classifications performed on the phone (watch+phone; *avg.*= 22.3%, *SD* = 11.5%, *max* = 42.3%) or on the cloud (watch+phone+cloud; avg. = 23.0%, *SD* = 10.8%, *max* = 39.8%). In these architectures, the watch was used only for recording sounds and supporting user interactions. For watch+cloud, the watch additionally computed the sound features and communicated with the cloud over WiFi, which resulted in significantly higher CPU utilization (*avg.*= 51.1%, *SD* = 14.9%, *max* = 76.1%). Finally, for the watch-only design, CPU utilization nearly maxed out (*avg.*= 99.0%, *SD* = 2.1%, *max* = 100%) because the classification model ran directly on the watch, revealing that this design is impractical for real-world use. However, future advancements in machine learning and wearable technology may

Figure 5. Average CPU (a) and memory (b) usage of the four architectures. Error bars show 95% confidence intervals.
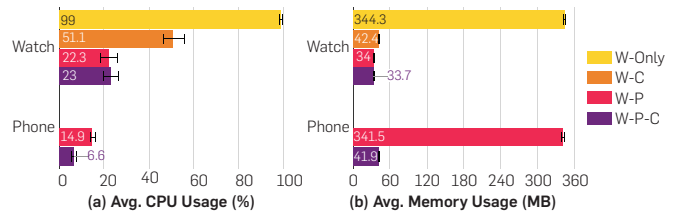
lead to smaller models and more powerful watches that can run these models locally.

Memory usage. A smartwatch app must be memory efficient. Unsurprisingly, we found that the memory usage heavily depended on where the model (281.8MB) was running; hence, watch-only and watch+phone consumed the highest memory on the watch (*avg.* = 344.3MB, *SD* = 2.3MB, *max* = 346.1MB) and phone (*avg.*= 341.5MB, *SD* = 3.0MB, *max* = 344.1MB), respectively (see Figure 5b). This indicates that running a large model such as VGG-lite on the watch will likely exceed the memory capacity of current smartwatches. The other app processes (e.g., UI and computing features) required less than 50MB of memory.

Network usage. Low network usage increases the app portability, especially in low-signal areas, and may help reduce Internet costs. Only the cloud-based architectures required network because the classifications were performed locally for watch- or phone-based designs. Specifically, for watch+cloud, the average network consumption, when the system was actively classifying sounds every second, was 486.8B/s (*SD* = 0.5B/s, *max* = 487.6B/s), and for watch+phone+cloud, it was 486.5B/s (*SD* = 0.5B/s, *max* = 487.2B/s), which is very low (∼1.8MB/h). In reality, sounds will likely not occur every second, which will reduce the total consumption even further.

Battery consumption. We measured the battery drain time from full charge until 30% (see Figure 6), finding that the watch-only architecture consumed a lot of battery: it reached 30% battery in 3.3 h only. Within the remaining architectures, both watch+phone (30% at 15.2 h) and watch+phone+cloud (30% at 16.1 h) were more efficient than watch+cloud (30% at 12.5 h), because the latter used WiFi, which consumes more energy than BLE.[19] Similar trends were observed on the phone; however, running the model on the phone (watch+phone) was still tolerable compared to the watch (see Figure 6). In summary, we expect the watch-only design would be impractical for daily use, whereas others are usable with the on-device implementations fairing slightly better than the cloud ones.

End-to-end latency. A real-time sound awareness system needs to be performant. Figure 7 shows a computational breakdown of end-to-end latency, that is, the total time taken in obtaining a notification for a produced sound. On average, watch+phone+cloud performed the fastest (*avg. latency* = 1.8 s, *SD* = 0.2 s) followed by watch+phone (*avg.*= 2.2 s, *SD* = 0.1 s), which needed more time for running the model on the phone (vs. cloud), and watch+cloud (*avg.*= 2.4 s, *SD* = 0.0 s), which required more time to compute features on the watch (vs. phone in watch+phone+cloud). As expected, watch-only

Figure 6. Battery level over time on the (a) watch and the (b) phone for the four architectures: watch only, watch+cloud, watch+phone, and watch+phone+cloud. Baseline represents the case without the SoundWatch app running.
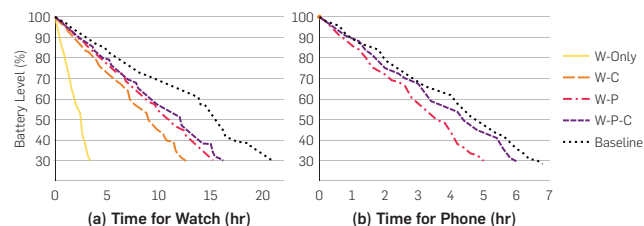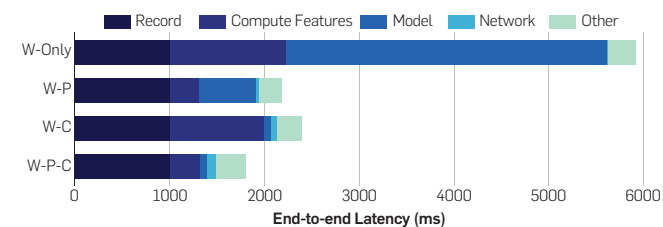

Figure 7. Breakdown of end-to-end latency for the four architectures.

was considerably slower (*avg.* = 5.9 s, *SD* = 0.1 s) and is, thus, currently unusable (though future smartwatches could be more capable). In summary, except for watch-only, all architectures had a latency of about 2 s; we evaluate whether this is acceptable in our user study.

Architecture evaluation summary. In summary, watch+phone and watch+phone+cloud outperformed the watch+cloud architecture for all system parameters. Additionally, the watch-only architecture was deemed impractical for real-life use due to high CPU, memory, and battery usage, and a large end-to-end latency. Among the phone-based architectures, watch+phone+cloud performed better than the watch+phone.

5. USER STUDY

To study end-user perceptions of our system results and reactions to SoundWatch across multiple contexts, we performed a lab and campus walkthrough evaluation with eight DHH participants. Although SoundWatch can support any architecture or model, we used only the best-performing architecture (watch+phone) and model (VGG-lite) for this study.

5.1. Participants

We recruited eight DHH participants (three women, three men, and two nonbinary) using email, social media, and snowball sampling. Participants were on average 34.8 years old (*SD* = 16.8, *range* = 20–63). Four had profound hearing loss, three had severe, and one had moderate. Seven reported onset as congenital and one reported one year of age. Seven participants used hearing devices: three used cochlear implants, one used hearing aids, and three used both. For communication, five participants preferred sign language and three preferred to speak verbally. All participants reported fluency with reading English (5/5 on rating scale, 5 is best). Participants received $40 as compensation.

5.2. Procedure

The in-person procedure took place on a university campus and lasted up to 90 min. Sessions were led by the lead author who is hard of hearing and knows level-2 ASL. A real-time transcriptionist attended all sessions and five participants opted to additionally have a sign language interpreter present. Questions were presented visually on an iPad, whereas responses and follow-up discussion were spoken or translated to/from ASL. The session began with a demographic questionnaire, followed by a three-part protocol:

Part 1: Introducing SoundWatch (5–10 min). First, we asked about general thoughts on using smartwatches for sound awareness. The researcher then demonstrated SoundWatch by explaining the UI and asking participants to wear the watch while making three example sounds (speech, door knock, and phone ring). Participants could also make their own sounds (e.g., by knocking or speaking).

Part 2: Campus walk (20–25 min). Next, the researcher and the participant (with the watch and phone) visited three locations on campus in a randomized order: (1) a home-like location (a building lounge), (2) an office-like location (a grad student office), and (3) an outdoor location (a bus stop). These locations allowed participants to experience Sound-Watch in different auditory contexts. In each location, participants used the watch naturally (e.g., by sitting on a chair in an office) for about 5 min. In locations with insufficient sound activity (e.g., if the lounge was empty), the researcher produced some sounds (e.g., by washing hands or opening a door). Before exiting each location, participants completed a short feedback form.

Part 3: Posttrial interview (45–50 min). After the campus walk, participants returned to the lab for a semistructured interview about their overall experience, perceptions of SoundWatch across the three locations, reactions to the UI, and any privacy concerns. We then asked about specific technical considerations, such as accuracy–latency trade-offs and the four possible SoundWatch architectures. For accuracy–latency, we gathered their expectations for minimum accuracy and maximum delay and whether these perspectives changed based on sound type (e.g., urgent vs. nonurgent sounds) or context (e.g., home vs. office). To help discuss the four SoundWatch architectures—and to more easily allow our participants to understand and track differences—we prepared a chart enumerating key characteristics such as battery or network usage with a HIGH, MEDIUM, or LOW rating based on our system experiment findings. Finally, we asked participants to rate the "ease-of-use" of each architecture (high, med, or low) by weighing factors such as the Internet requirement, number of devices to carry (e.g., 1 for watch-only vs. 2 for watch+phone), and the size of visual display (e.g., small for watch vs. medium for phone) and provide reasoning for their choice.

5.3. Data analysis

We analyzed the interview transcripts and the in situ form responses using an iterative coding approach.[2] To begin, we randomly selected three out of eight transcripts; two researchers read these transcripts and developed an initial

codebook. The researchers then independently assigned codes to the three transcripts, while simultaneously refining their own copy of the codebook (adding, merging, or deleting codes). The researchers then met again to discuss and refine the codebook, resulting in 12 level-1 codes and 41 level-2 codes) arranged in a hierarchy. This final codebook was used to code the remaining five transcripts by the two coders, resulting in an interrater agreement (measured using Krippendorff's alpha) of 0.79 ($SD = 0.14$, $range = 0.62$-1) and a raw agreement of 93.8% ($SD = 6.1\%$, $range = 84.4\%$-100). Conflicting code assignments were resolved via consensus.

5.4. Findings
We detail participants' experience with SoundWatch during the campus walk as well as comments on model accuracy–latency, system architectures, and the user interface.

Experience with campus walk. All participants found the watch generally useful to help with the everyday activities in all three contexts (home-like lounge, office, and outdoors). For example,

> "My wife and I tend to leave the water running all the time so this app could be beneficial and save on water bills. It was helpful to know when the microwave beeps instead of having to stare at the time [microwave display]." (P6)
> "This is very useful for desk type work situations. I can use the watch to help alert me if someone is knocking the door, or coming into the room from behind me." (P7)

However, all participants also reported problems, the most notable being delay and misclassifications; the latter were higher in outdoor contexts than in others. For example,

> "The app is perfect for quiet settings such as [at] home. [While outdoors,] some sounds were misinterpreted, such as cars were recognized as water running." (P3)

In situ feedback form responses corroborate these comments, with average usefulness for lounge (4.8/5 on a rating scale (5 is best), $SD = 0.4$) and office (4.6/5, $SD = 0.5$) being higher than for outdoors (3.5/5, $SD = 0.5$).

Even with a low usefulness rating in outdoor settings, all participants wanted to use the app outdoors, mentioning that they may be able to use contextual information to supplement inaccurate feedback. For example,

> "Sure there were some errors outdoors, but it tells me sounds are happening that I might need to be aware of, so I can look around and check my environment for cues." (P8)

Model accuracy–latency comparison. Deep learning-based sound recognition will never be 100% accurate. Thus, we asked participants about the minimum required accuracy and the maximum tolerable delay at which they will use a smartwatch app. The most common preference was a maximum delay of "five seconds" (5/8) and a minimum accuracy of 80% (6/8); however, this choice was additionally modulated by the specific sound type. Specifically, for urgent sounds (e.g., fire alarms or car horn), participants wanted the minimum possible delay (at the cost of accuracy) to get quick information for any required action, because "I'll at least know something is happening around me and [...] can look around to see if a car is honking" (P2).

In contrast, for nonurgent sounds (e.g., speech and laughing), more accuracy was preferred because participants mentioned that repeated errors could be annoying (7/8). For example:

> "I don't care about speech much, so if there is a conversation, well fine, doesn't matter if I know about it 1-2 second later or 5 seconds later, does it? But if it makes mistakes and I have to get up and check who is speaking every time it makes a mistake, that can be really frustrating." (P5)

Finally, for medium-priority sounds (e.g., microwave for P3), participants (7/8) wanted a balance, tolerating a moderate amount of delay for moderate accuracy.

Besides sound type, preference also varied with the context of use (home vs. office vs. outdoors). Participants preferred having less delay in more urgent contexts and vice versa. *That is,* for the home, participants (8/8) wanted high accuracy—and accepted more delay—because, for example:

> "I know most of what is going on around my home. And at home, I am generally more relaxed [so] delay is okay. But, I don't want to be annoyed by errors in my off time." (P8)

For the office, participants (6/8) felt they would tolerate a moderate level of accuracy with a moderate level of delay, because *"something may be needing my attention but it's likely not a safety concern to be quick about it"* (P8). Preferences for outdoors were split: four participants wanted a minimum delay (at the cost of accuracy), but the other four did not settle for a single response, mentioning that the trade-off would depend on the urgency of the specific sound:

> "If it's just a vehicle running on the road while I am walking on the sidewalk, then I would want it to only tell if it's sure that it's a vehicle running, but if a car is honking say if it behind me, I would want to know immediately." (P2)

Architecture comparison. By saliently introducing the performance metrics (e.g., battery usage) and usage requirements (e.g., Internet connection for cloud), we gathered qualitative preferences for the four possible SoundWatch architectures: watch-only, watch+phone, watch+cloud, and watch+phone+cloud.

In general, watch+phone was the most preferred architecture among all participants, because, compared to watch-only, it is faster, requires less battery, and has more visual state available for customization. In addition, compared to cloud-based designs, watch+phone is more private and self-contained (does not need Internet).

However, five participants wanted the option to be able to customize the architecture on the go, mentioning that in outdoor settings, they would instead prefer to use watch+phone+cloud because of speed and accuracy advantages. This is because in the outdoor context, data privacy is of less concern for them. For example:

> "Accuracy problems could be more [outdoors] due to background noise and [thus] I prefer to use cloud for [stronger] models if [the] internet is available. At home/office, there is a possibility of private data breach." (P6)

Watch+cloud was preferred by two participants only for cases where it is hard to carry a phone, such as in a *"gym or [while] running outdoors"* (P1). Finally, watch-only was not preferred for any situation because of high battery drain.

User interface suggestions. Overall, participants appreciated the minimalistic app design and the customization options (mute button and checklist on phone). When asked about future improvements, they suggested three: (1) show the urgency of sounds—for example, using vibration patterns or visual colors; (2) show direction of sounds, particularly for outdoor contexts; and (3) explore showing multiple sounds to compensate for inaccuracy:

> *"You could give suggestions for what else sound could be when it's not able to recognize. For example, [...] if it is not able to tell between a microwave and a dishwasher, it could say "microwave or dishwasher", or at least give me an indication of how it sounds like, you know like a fan or something, so I can see and tell, oh yeah, the dishwasher is running." (P4)*

6. DISCUSSION
Our work reaffirms DHH users' needs and preferences for smartwatch-based sound awareness[5, 17] but also (1) implements and empirically compares state-of-the-art deep learning approaches for sound classification on smartwatches, (2) contributes a new smartwatch-based sound identification system with support for multiple device architectures, and (3) highlights DHH users' reactions to accuracy–latency trade-offs, device architectures, and potential concerns. Here, we reflect on further implications and limitations of our work.

6.1. Utility of sound recognition
How well does sound recognition tool need to perform to provide value? Our findings show that this is a complex question that requires further study. Although improving overall accuracy, reducing latency, and supporting a broad range of sound classes is clearly important, participants felt that urgent sounds should be prioritized. Thus, we wonder, would an initial sound awareness app that supports three to ten urgent sounds be useful? One way to explore this question is to study SoundWatch—or a similar app—over a longitudinal period with multiple customization options. However, this approach also introduces ethical and safety concerns as automatic sound classification will never be 100% accurate. High accuracy on a limited set of sounds could (incorrectly) gain the user's trust, and the app's failure to recognize a safety sound (e.g., a fire alarm) even once could be dangerous. In general, a key finding of our research is that users desire customization (e.g., which sounds to classify) and transparency (e.g., classification confidence).

6.2. Toward improving accuracy
Our user study suggests a need to further improve system accuracy or at least explore other ways to mitigate misclassification. One possibility, suggested by P4, is to explore showing multiple "possible" sounds instead of the most probable sound—just as text autocomplete shows n-best words. Another idea is to sequentially cascade two models, using the faster model to classify a small set of urgent sounds and the slower model for

lower-confidence classifications and less-urgent sounds. End-user customization should also be examined. Each user could select the desired sounds and the required accuracy, and the app could dynamically fine-tune the model (e.g., by using a weighted class accuracy metric). Finally, as proposed by Bragg et al.,[1] researchers should explore end user interactive training of the model. Here, guided by the app, participants could record sounds of interest to either improve existing sound classes or to add new ones. Of course, this training may be tedious and difficult if the sound itself is inaccessible to a DHH user.

6.3. Privacy implications
Our participants were concerned with how cloud-based classification architectures may invade their own "sound" privacy and of others around them. However, uploading and storing data on the cloud have benefits. These datasets can be used for improving the classification model. Indeed, modern sound architectures on IoT devices (e.g., Alexa and Siri) use the cloud for exchanging valuable data. A key difference to our approach is that these devices only transmit after listening to a trigger word. Thus, what are the implications for future always-listening sound awareness devices? We see three. First, the users should have control over what data gets uploaded, which can be customized based on context (e.g., offices might have more private conversations than outdoors). Second, future apps will need clear privacy policies such as GDPR or CCPA that outline how and where the data is stored and what guarantees the users have. Finally, users should always have access to their data with an option to potentially delete it, in entirety, from the cloud.

6.4. Future smartwatch applications
In contrast to past wearable sound awareness work,[6,9,13] we used commercially available smartwatches, a mainstream popular device that is more socially acceptable than HMDs[6,9] or custom hardware-based[6] solutions—and that may be preferred over smartphones for sound recognition feedback.[3] So, what are other compelling applications of a smartwatch-based sound awareness for DHH users? Full speech transcription, a highly desired feature by DHH users,[3] is difficult to accommodate on the small watch screen, but future work could explore highlighting important keywords or summarizing key conversation topics. Sound localization is also desired[1,5] and could be investigated by coupling the watch with a small external microphone array or designing a custom watch with multiple microphones. However, how best to combine different sound and speech features (e.g., topic summary, direction, and identity) on the watch is an open question. Goodman et al.[5] investigated designs for combining sound identity, direction, and loudness on watch; however, this study was formative with a focus on user interface. Future work should also explore the system design aspects of showing multiple features—a challenging problem given the smartwatch's low-resource constraints.

6.5. Limitations
First, although our sound recognition technology is heavily informed by DHH perspectives, such as those of our

hard-of-hearing lead author, we do not assume it is universally desired. Some DHH people may feel negatively toward this technology, especially those who identify as part of deaf culture.[1,3] At the same time, past work,[1,3] such as our own survey with 201 DHH participants,[3] suggests the DHH community is broad and many DHH individuals do find sound recognition valuable. Still, future work should continue to examine preferences for sound feedback with a diverse section of the DHH population to verify our findings.

Second, our short 20-min campus walk, although useful as an initial, exploratory study, could not investigate pragmatic issues, such as user perception of battery life and long-term usage patterns. Future work should perform a longitudinal deployment and compare results with our lab findings.

Third, our model accuracy results, though gathered on real-life recordings of 20 sounds, do not accurately reflect real-world use where other sounds beyond those 20 could also occur. Although our approach provides a baseline for model comparison and contextualizing user study findings, a more accurate experiment would include a post hoc analysis of sound data collected from longitudinal watch use.

Finally, we evaluated our models on specific hardware devices (Ticwatch Pro Watch, Honor 7x Phone). Although the relative comparisons are likely generalizable, the absolute performance metrics will change as the mobile and wearable technologies evolve in the future. Additional studies will be needed then.

7. CONCLUSION
In this paper, we performed a quantitative examination of modern deep learning-based sound classification models and architectures as well as a lab exploration of a novel smartwatch sound awareness app with eight DHH participants. We found our best classification model performed similar to the state of the art for nonportable devices although requiring a substantially less memory (∼1/3rd) and that the phone-based architectures outperformed the watch-centric designs in terms of CPU, memory, battery usage, and end-to-end latency. Qualitative findings from the user study contextualized our system experiment results and uncovered ideas, concerns, and design suggestions for future wearable sound awareness technology.

Acknowledgments

References
1. Bragg, D., Huynh, N., Ladner, R.E. A personalizable mobile sound detector app design for deaf and hard-of-hearing users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (2016), ACM Press, New York, 3–13.
2. Braun, V., Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol. 3*, 2 (2006), 77–101.
3. Findlater, L., Chinh, B., Jain, D., Froehlich, J., Kushalnagar, R., Lin, A.C. Deaf and hard-of-hearing individuals' preferences for wearable and mobile sound awareness technologies. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2019), ACM, Glasgow, UK, 1–13.
4. Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M. Reliable detection of audio events in highly noisy environments. *Pattern Recognit. Lett. 65*, (2015), 22–28.
5. Goodman, S., Kirchner, S., Guttman, R., Jain, D., Froehlich, J., Findlater, L. Evaluating smartwatch-based sound feedback for deaf and hard-of-hearing users across contexts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2020), ACM, Honolulu, Hawaii, 1–13.
6. Gorman, B.M. VisAural: A wearable sound-localisation device for people with impaired hearing. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility* (2014), ACM, Rochester, NY, 337–338.
7. Guo, R., Yang, Y., Kuang, J., Bin, X., Jain, D., Goodman, S., Findlater, L., Froehlich, J. Holosound: Combining speech and sound identification for deaf or hard of hearing users on a head-mounted display. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (2020), ACM, 1–4.
8. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2017), IEEE, New Orleans, LA, 131–135.
9. Jain, D., Findlater, L., Volger, C., Zotkin, D., Duraiswami, R., Froehlich, J. Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM CHI, Seoul, Korea, 241–250.
10. Jain, D., Lin, A.C., Amalachandran, M., Zeng, A., Guttman, R., Findlater, L., Froehlich, J. Exploring sound awareness in the home for people who are deaf or hard of hearing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, Glasgow, UK, 94:1–94:13.
11. Jain, D., Mack, K., Amrous, A., Wright, M., Goodman, S., Findlater, L., Froehlich, J. HomeSound: An iterative field deployment of an in-home sound awareness system for deaf or hard of hearing users. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20 (New York, NY, USA, 2020), Association for Computing Machinery, Honolulu, Hawaii, 1–12.
12. Jain, D., Ngo, H., Patel, P., Goodman, S., Findlater, L., Froehlich, J. SoundWatch: Exploring smartwatch-based deep learning approaches to support sound awareness for deaf and hard of hearing users. In *ACM SIGACCESS Conference on Computers and Accessibility* (2020), ACM, 1–13.
13. Kaneko, Y., Chung, I., Suzuki, K. Light-emitting device for supporting auditory awareness of hearing-impaired people during group conversations. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference* (2013), IEEE, Manchester, UK, 3567–3572.
14. Lu, L., Zhang, H.-J., Jiang, H. Content analysis for audio classification and segmentation. *IEEE Trans. Speech and Audio Process. 10*, 7 (2002), 504–516.
15. Matthews, T., Fong, J., Ho-Ching, F.W.-L., Mankoff, J. Evaluating non-speech sound visualizations for the deaf. *Behav. Inf. Technol. 25*, 4 (2006), 333–351.
16. Mazumdar, A., Haynes, B., Balazinska, M., Ceze, L., Cheung, A., Oskin, M. Perceptual compression for video storage and processing systems. In *Proceedings of the ACM Symposium on Cloud Computing* (2019), ACM, Santa Cruz, CA, 179–192.
17. Mielke, M., Brück, R. A pilot study about the smartwatch as assistive device for deaf people. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (2015), ACM, Lisbon, Portugal, 301–302.
18. Saunders, J. Real-time discrimination of broadcast speech/music. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (1996), Vol. 2, IEEE, Atlanta, GA, 993–996.
19. Shahzad, K., Oelmann, B. A comparative study of in-sensor processing vs. raw data transmission using ZigBee, BLE and Wi-Fi for data intensive monitoring applications. In *2014 11th International Symposium on Wireless Communications Systems (ISWCS)* (2014), IEEE, Barcelona, Spain, 519–524.
20. Sicong, L., Zimu, Z., Junzhao, D., Longfei, S., Han, J., Wang, X. UbiEar: Bringing location-independent sound awareness to the hard-of-hearing people with Smartphones. *Proc. ACM on Interact. Mob. Wearable and Ubiquitous Technol. 1*, 2 (2017), 17.
21. Yeung, E., Boothroyd, A., Redmond, C. A wearable multichannel tactile display of voice fundamental frequency. *Ear Hear. 9*, 6 (1988), 342–350.
22. Yuan, H., Reed, C.M., Durlach, N.I. Tactual display of consonant voicing as a supplement to lipreading. *J. Acoust. Soc. Am. 118*, 2 (2005), 1003.

Dhruv Jain, Hung Ngo, Pratyush Patel, Khoa Nguyen, and Jon Froehlich ({djain, hvn297, patelp1, akhoa99, jfroehli}@uw.edu), Computer Science and Engineering, University of Washington, Seattle, WA, USA.

Steven Goodman, Rachel Grossman-Kahn, and Leah Findlater ({smgoodmn, rachelgk, leahkf}@uw.edu), Human Centered Design and Engineering, University of Washington, Seattle, WA, USA.