Sound Sensing and Feedback Techniques for Deaf and Hard of Hearing People

Dhruv Jain

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jon E. Froehlich, Chair

Leah Findlater, Chair

Richard Ladner

Program Authorized to Offer Degree:

Computer Science and Engineering

Dhruv Jain

University of Washington

University of Washington

Abstract

Sound Sensing and Feedback Techniques for Deaf and Hard of Hearing People

Dhruv Jain

Chair of the Supervisory Committee:

Jon E. Froehlich
Computer Science and Engineering

Leah Findlater
Human Centered Design and Engineering

Sound awareness can have wide ranging impact for people who are deaf or hard of hearing (DHH) from being informed of safety critical sounds such as fire alarms and sirens to more mundane but useful cues such as microwave beeps or door knocks. Unfortunately, the current solutions used by DHH people—such as flashing doorbells or vibratory alarm clocks—only substitute for some specific sounds, but do not provide general sound awareness. In this dissertation, I report on several iterative studies aimed at: understanding sound awareness needs and preferences of DHH people in multiple contexts, designing systems to provide sound and speech awareness in those contexts, and evaluating these systems in both controlled environments and in the field. I close with outlining the future grand challenges in the emerging field of sound accessibility.

Full Committee

Jon E. Froehlich, Chair

Leah Findlater, Chair

Richard Ladner

Jennifer Mankoff

Jeffrey Bigham

Jacob O. Wobbrock, GSR

Table of Contents

Acknowledgements

Dissertations contain one name, but mine was a product of intense collaboration and support from many people. First off are my advisors, Jon Froehlich and Leah Findlater, who I owe immense gratitude for every word in this document—thank you for your monumental hands-on guidance! While I cannot ever compensate you for your efforts, I will carry the torch of your hard work as best as I can during my faculty career.

Next are my research co-authors, namely, Steven Goodman, Emma McDonnell, Kelly Mack, Venkatesh Potluri, Pratyush Patel, Aditya Kusupati, Ather Sharif, Rose Guttman, Bonnie Chinh, Rachel Franz, Jackson Cannon, Akli Amrous, Matt Wright, Ana Liu, Aileen Zeng, Marcus Amalachandran, Angela Lin, Greg Guo, Robin Yang, Hung Ngo, Khoa Nguyen, Quan Dang, Hang Do, Rachel Grossman-Kahn, and Raja Kushalnagar whose grind and intellect form the backbone of my dissertation research—thank you! I also owe my colleagues and friends, especially, Liang He, Manaswi Saha, Venkatesh Potluri, Pratyush Patel, and Sudheesh Singanamalla who tolerated my vigorous energy and patiently listened and provided feedback during the most needed times—thank you all!

My committee members, Richard Ladner, Jennifer Mankoff, Jeffrey Bigham, and Jacob Wobbrock supported me not only in completing this dissertation, but also throughout my professional career. I have known and consulted my committee members since my undergraduate years; it means a lot that their doors have always been open for me for any professional and personal advice—thank you really!

I especially wanted to thank and congratulate the support staff at CSE UW, without whom no work would have been possible. These geniuses provide infrastructure and moral support for every research and personal development, yet they often do not appear on our research papers. At CSE, every staff member has been extremely responsive, extremely diligent, extremely helpful, immensely knowledgeable, and very kind. I particularly interacted with support staff a lot during my graduate career, including with Elise Dorough, Chiemi Yamaoka, Aaron Timss, Emma Gebben, Tracy Erbeck, Eric Eto, Joe Eckert, Hector Rodriguez, Rebekah Hanson, Emma Notkin, CJ Smith, Sandy Kaplan, David Kessler, Kristin Osborne, Kay Beck-Benton, Elle Brown, Amber Cochran, Sophie Ostlund, Adrian Dela Cruz, Leslie Sessoms, John Akers, and

the former and new chairs, Hank Levy and Magdalena Balazinska. I will shout your names from the rooftops—you all are really awesome—thank you, thank you!

Being an international student, I was away from home a lot during my PhD. I was extremely lucky to have support of my family members back home, especially my dad, Alok Jain, mom, Ruchi Gupta, and brother, Kunal Jain as well as my cousins, uncles, aunts, and grandparents who have been rooting for me from across the world (and the late ones, from above)! I have a big family and every single one of them have cheered for me all throughout. They read all my papers, celebrated every achievement, and consoled me in every downtime. I cannot stress this enough—I am extremely blessed to have you! You have made my PhD journey and life immensely satisfying.

And finally, most importantly, …

To Tanu, the hidden force behind all my work.

I could not have done this without you.

Chapter 1: Introduction

The world is filled with a rich diversity of sounds such as dog barking, door knocking, or a microwave beeping. These sounds help provide important information to perform daily tasks (*e.g.*, getting food from a microwave), get away from an impending danger (*e.g.*, by evacuating a building with a fire alarm), or feel present in nature (*e.g.*, by hearing a bird chirp). These sounds, however, can be inaccessible to a large section of the population who have trouble hearing. The National Institute of Deafness and Other Communication Disorders (NIDCD), a subsidiary of National Institute of Health (NIH) reports that over 15% of US adults have some trouble hearing [144]. The rates of disabling hearing loss are lower, but they rise substantially with age from 2% at 45-54 years of age to over 50% of those who are 75 and older.

Note that not all d/Deaf and hard of hearing (DHH) people want to be aware of sounds. Many of them use alternative ways to deal with information typically conveyed to hearing people through sounds such as flashing light-based doorbells or vibratory alarm clocks. At the same time, large scale surveys with DHH people [13,25] have shown that many of them are interested in having greater access to speech and sound information. Indeed, this unmet need is evident though continuing widespread adoption of sound awareness technologies such as hearing aids, cochlear implants, and more recently, Google's *Live Transcribe* [145]. While these technologies help improve sound and speech recognition to some extent, they only provide awareness about some specific types of sounds, do not work in a variety of contexts, and are not always comfortable to wear [66,83]. Our recent survey with 201 DHH participants [25] showed that, even after using their assistive technologies, 73.1% of the participants were still "extremely" or "very interested" in having greater access to sounds around them.

Informed by the lived experiences of DHH people—including my own as a hard of hearing individual [48,57]—I am exploring *new sound awareness approaches for DHH people* that leverage advances in sound processing, machine learning, and augmented reality to sense and provide sound feedback. In my dissertation, I explored four related threads of work:

1. Visualizing sounds in the home using stationary displays
2. Providing sound awareness on portable devices
3. Personalizing sound feedback on portable devices
4. Improving speech awareness using head-mounted displays

1.1 Thesis Statement

My proposed thesis was:

> *"For DHH people who want sound feedback, providing this feedback unobtrusively and glanceably will enhance their understanding of sounds and information conveyed by sounds."*

To explore this thesis, I followed an iterative human-centered process in my research ranging from formative studies to design and evaluation of prototypes in controlled environments to deployments of full systems in the field (Figure 1.1).

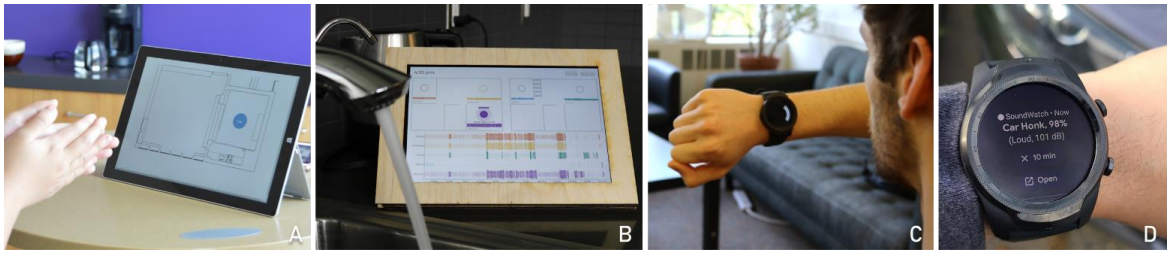


Figure 1.1: My research explores sound accessibility for DHH users through an iterative design process from large-scale surveys (not shown) to controlled evaluations of system-designs (A and C) to full system deployments in the field (B and D).

1.2 Document Outline and Organization

As outlined above, I explore four threads of work in this dissertation: visualizing sounds in the home, providing sound awareness on portable devices, personalizing sound feedback, and improving speech awareness through head-mounted displays.

Visualizing sounds in the home. In the first exploration, outlined in Chapter 3, I address *how to classify and visualize sounds in the home*. Commonly used commercial products—such as flashing doorbells or vibratory alarm clocks—substitute specific auditory cues with visual or haptic feedback, but do not provide general awareness about sounds in the home.

In response, through an iterative design process (Figure 1.1a, 1.1b) [53], I created *HomeSound* [54] an IoT based system visualizes sound activity in the whole house. HomeSound contains a network of interconnected "picture-framed" displays running a deep-learning engine deployed in different rooms of the home (Figure 1.1b). Each display senses and classifies 19 common sounds in the home (*e.g.*, dog bark, doorbell) and interacts with a centralized server to produce a single across-home visualization of in-home sound activity.

I iteratively deployed HomeSound in total six homes of DHH people, starting with visualizations of simple but accurate sound characteristics (*e.g.*, loudness, pitch), before deploying more complex but error-prone sound information (*e.g.*, the sound type). After a three-week use, DHH participants reported being able to perform some important daily tasks more easily and quickly (*e.g.*, getting clothes from a dryer, turning off the kitchen fan). For some users, HomeSound fundamentally changed their understanding of sounds in their home. For example,

> *"I became aware of so many sounds in my home: dryer whirring, cutlery clinking, door opening, water running. My wooden home makes a lot of noise when I walk. I didn't realize we have a noisy home. Do we make more noise than hearing people?!"*

We also uncovered privacy concerns with our always-recording setup in a personal context of the "home". Overall, our findings provide important insights to design AI-based sound sensing systems accounting for concerns of social dynamics, space, privacy, and trust.

Providing sound awareness on portable devices. As the second key area, I explored *how to provide accurate sound awareness in portable environments* (Chapter 4). After examining several portable devices, I chose to use smartwatch, the most preferred device by DHH people in our survey [25], since it is private, glanceable, and always-available on the wrist.

However, classifying sounds on a low-resource device and displaying sound information on a small watch screen is a challenging problem. To address, our team performed two smartwatch-based studies. First, in a *Wizard-of-Oz* evaluation, led by my colleague Steven Goodman, we explored several dense sound feedback designs [33], identifying preferences for visualizations and complementary vibratory patterns to accompany the visualizations—another benefit of smartwatch. Second, using transfer learning, I trained and compared

four small deep-learning models to accurately classify sounds on smartwatches [55], identifying a promising model that performed close to state-of-the-art for non-portable devices (82% accuracy), while requiring substantially less memory and computational power.

Based on the above insights, I built *SoundWatch*, a real-time sound recognition app for commodity Android smartwatches [55]. SoundWatch uses our privacy-preserving sound recognition pipeline and can be switched among four system architectures (*watch-only*, *watch+phone*, *watch+cloud*, *watch+phone+cloud*). Through a user evaluation with eight DHH users, I uncovered several implications for future wearable accessibility systems such as the need for 'switchable' system architectures, multiple cascaded models, end-user control of the data, and a customizable user interface.

After improvements, we released the app with our best performing model and system architecture on Google Play Store; see a demo video. Thus far, it has been downloaded over 2000 times. This line of work has received extensive press coverage, a best artifact award at ASSETS 2020, and was invited for CACM Research Highlights, which publishes the most significant recent research results across the field of computing.

Personalizing sound feedback. A key desired feature by DHH people in our survey [25] and system evaluations [54,55] was end-user personalization—such as training on new sound categories (*e.g.*, a new custom home appliance) or users' specific sounds (*e.g.*, my child's voice) [25,54,55]. However, prior sound recognition systems, including HomeSound and SoundWatch noted above, typically require a large amount of training data, making it difficult personalize the model *in-situ* for custom sounds.

In Chapter 5, I present the design and development of *ProtoSound* [56], an interactive system that allows users to personalize a sound recognition engine by recording only a few training samples (*e.g.*, five for each sound). The key idea is that by training a model repeatedly over datasets of limited training samples, it learns to train rapidly from a few samples in the field.

To evaluate ProtoSound, I quantified performance on real-life sound datasets I compiled [54,55] and conducted a field deployment study with 19 participants. Results show that ProtoSound significantly

outperformed the best baseline (9.7% accuracy advantage), required minimal end-user effort to train (~10 minutes), and accurately learned sounds across many real-world locations (*e.g.*, homes, restaurants, grocery stores, and parks).

Improving speech awareness through augmented reality. I also examined *how to support speech conversations for DHH users* using AR-style head-mounted displays (HMD) (Chapter 6), which offer benefits of glanceability and reduced visual split over the traditional stationary captioning approaches. Again, an iterative design and evaluation process resulted in several prototypes, starting with an initial captioning prototype [46], to a more refined prototype [50], and culminating in a full system [37] that displays multiple cues on an HMD (speech, location, non-speech sounds) (Figure 1.2). The final system is open-sourced; see a demo video.

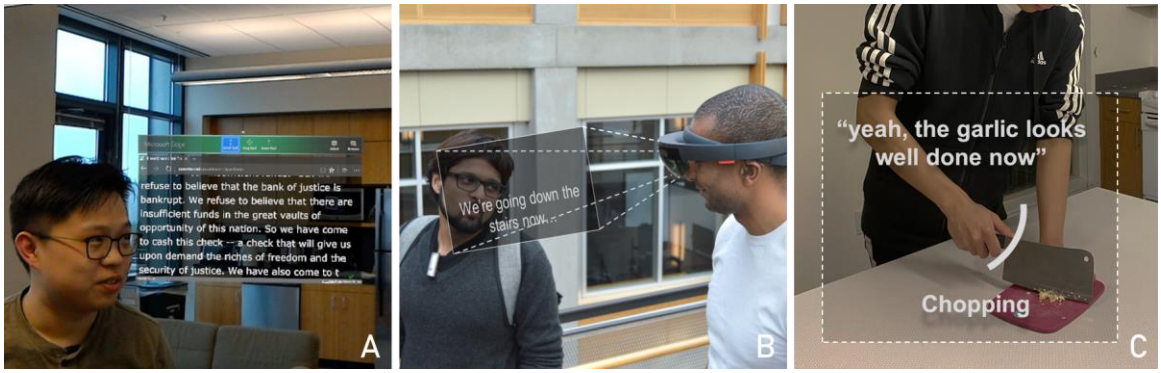


Figure 1.2: Iterative exploration of speech awareness on heads-mounted display (HMD) starting from an initial captioning prototype (A) to a more refined prototype (B) and, finally, a full system that combines captions, location, and sound recognition cues on an HMD (C).

1.3 Summary of Contributions and Broad Impacts

My dissertation work makes contributions in human-computer interaction literature through eight scholarly publications, including in CHI, ASSETS, and DIS. These contributions span the full human-centered research pipeline and include: (1) uncovering sound and speech awareness preferences of DHH people through formative studies, (2) design of system prototypes to provide speech and sound awareness feedback to DHH users, and (3) insights from evaluation of these prototypes in multiple contexts (at homes, while mobile, outdoors, during conversations), including design guidelines for future sound and speech awareness technologies.

Beyond intellectual contributions, my research has resulted in (and has the potential for continued) real-world impact. All codebase for my work is open sourced for researchers to build upon. Our SoundWatch app is launched and freely available on Google Play Store (over 2000 downloads so far), which we continue to improve based on feedback from our users. As well, sound recognition feature is now integrated, by default, on iPhone and Android smartphones.

Such advances have the potential to impact a large portion of the population. Approximately 15% of US adults report some trouble hearing; those with "disabling hearing loss" account for 2% of adults aged 45 to 54, a number that increases substantially with age to nearly 50% of those 75 and older [93]. Beyond DHH users, sound feedback may also be desired by hearing people or those with cognitive disabilities, particularly in cases of situational impairments [118] (*e.g.*, when wearing headphones) or cognitive overload (*e.g.,* when facing too much noise). Finally, the ability to sense and unobtrusively display sound information has applications in many other domains, such as wildlife surveys, home automation, and appliance repairs.

1.4 Authorship Statement

Although I am the principal author of the research detailed in this dissertation, it is also the product of years of collaboration with my advisors, Jon Froehlich and Leah Findlater, and my colleagues, including Steven Goodman, Emma McDonnell, Hung Ngo, Khoa Nguyen, and Rachel Grossman-Kahn, at the University of Washington. To acknowledge my collaborators' contributions, I use first-person plural in the following chapters.

Much of my work is also informed by prior publications from my colleagues, including a large-scale survey of DHH users' sound feedback preferences led by my advisor Leah Findlater [2], a formative evaluation of smartwatch based sound awareness designs led by my colleague Steven Goodman [3], and a AR speaker localization work, GlassEar [10], conducted during my summer internship with my advisors (and colleagues) at their former institute, University of Maryland.

Chapter 2: Background & Related Work

I provide a background on DHH culture and technology adoption and contextualize my dissertation work within sound awareness needs and tools of DHH people as well as prior sound recognition research.

2.1 DHH Culture and Technology Adoption

For many DHH people, the degree of hearing loss is only a small aspect of their disability and does not determine their language preference or choice of accessible solutions [15,67,91]. To understand what factors affect inclusion, researchers have composed three models of deafness: medical, social and cultural [15,143]. In the medical model, a person with hearing loss is seen as wanting to restore normal hearing. In the social model, a DHH individual is considered as needing to integrate into the society of hearing people. Finally, in the cultural model, a DHH person is viewed as part of a culture or community with a distinct visual language. Usage of these models depends on the research goals [15,67]. For example, to develop hearing aids and cochlear implants for (partially) restoring hearing, researchers primarily embody the medical model [15]. In my research, I adopt social and cultural models; I examine DHH individuals' interaction with people, culture, and environments to design accessible technologies.

Within the cultural model, an individual can identify as deaf, Deaf (capital 'D') or hard of hearing. The term Deaf refers to people who belong to a Deaf culture with common language, values and practices (see [15,71,91] for details). In contrast, the terms deaf and hard of hearing indicate someone for whom deafness is primarily an audiological experience and who refrain from membership to a particular community [15,91]. Individuals who identify as 'deaf' or 'hard of hearing' do not have a distinct cultural identity of their own, and they may choose to interact with either hearing or Deaf people based on their comfort [15,91].

These cultural differences may influence technology adoption. In the past, Deaf people have often criticized accessible technologies invented by hearing people as audist (i.e., conforming to behaviors and practices of hearing culture, see [15] for a historical perspective). Yet, others have increasingly adopted common accessible technologies such as hearing aids, cochlear implants, and speech transcribers [15,17,104]. While these tools may enhance sound awareness, they have also been criticized as visible signs of disability which

could hinder social acceptance of the users [97,118]. Communication partners may also incorrectly perceive these technologies as eliminating disability, thus making communication even more difficult [118]. Yet, some people, including me, view personal accessible technologies *"as expressions of identity, as fashion items and indicators of technical prowess"* [117]. Regardless of these opposing views, mainstreaming of some accessible technologies (*e.g.,* movie subtitles, noise cancelling earphones) have *"leveled the playing field"* between hearing and DHH people in some contexts [76]. My work builds on the "lived experiences" of many DHH individuals, including my own [48,57].

2.2 Sound Awareness Needs of DHH people

Designing effective sound awareness technology requires understanding the wide-ranging abilities and preferences of the DHH community. Several studies have surveyed DHH people on sounds of interest, highlighting a strong desire for urgent and safety-related sounds (*e.g.,* alarms, sirens) and social interaction support (*e.g.*, name calls, door knocks) [13,53,80,89,119]. Our recent survey of 201 DHH participants [25] further supported these findings, and showed that respondents were *"very"* or *"extremely"* interested in sound awareness, with particular value seen in using HMDs for captions and smartwatches for other types of sound notifications (*e.g.,* alarms, appliance sounds).

Sound interest is also influenced by cultural and contextual factors. For example, people who prefer communicating orally are more interested in sound and speech awareness than those who prefer sign language [25,49]. DHH respondents in the survey mentioned above [25] also predicted that social context (*e.g.,* with friends vs. strangers) would impact the use of a sound awareness tool, and a majority desired to have sound filtering rather than being informed of all sensed sounds.

Researchers have also studied what sound characteristics and feedback modalities are desired, finding that some (identity, location, urgency) are generally more important than others (volume, duration, pitch) [13,25] and that, for wearable devices, both visual and vibration feedback is needed. However, relative utility may differ by location or how the information is conveyed. For example, in the home, sound identity may be adequate [53], while directional indicators are important when mobile [89]. Similarly, for a smartwatch,

vibration is desired for notifying about sound occurrence with additional information supplemented through a visual display [13,53].

My work contributes to this literature by examining DHH people's sound awareness needs and technology preferences in multiple contexts (*e.g.,* at home, in transit, while mobile).

2.3 Non-Speech Sound Awareness Technologies

Commonly used assistive technologies by DHH people include hearing aids and cochlear implants. While these devices may improve sound recognition, they do not eliminate disability. Furthermore, several studies [66,83] also report low satisfaction with these aids due to issues with noise, comfort, and high cost. As a result, DHH people use other commercial products—such as flashing doorbells and vibrating wake-up alarms—as haptic or visual alternatives to some information typically conveyed by sound. While useful for their specific applications, these devices do not offer a general alternative to environmental sounds.

In the research literature, Matthews *et al.* [80] evaluated desktop-based prototypes for displaying sound visualizations in an *office* setting, identifying the preferred sound information to display (*e.g.*, sound source, location) and the visualizations for showing this information (*e.g.*, spectrograph, rings). In the HomeSound work [54], I instead investigate a IoT-based sound awareness system for the *home* and conduct field studies in six homes of DHH occupants.

In terms of portable solutions, researchers have explored smartphone [13,119], smartwatch [33,88], HMD [35,43,49], or wrist-worn displays [47,61,102] to convey sound identity (*e.g.,* a door knock), sound loudness, or source direction. For example, Bragg *et al.* [13] or Sicong *et al.* [119] and built smartphone apps to recognize and display environmental sounds (*e.g.,* phone ringing, sirens), conducting evaluations in the lab [13] or in a school [119]. DHH participants in both studies wanted both visual and vibration modalities to notify about sounds and a custom notification style for each specific sound (*e.g.,* using a different vibratory pattern). My SoundWatch work [55] builds on these findings to examine sound awareness on a smartwatch—the most preferred device for non-speech sound feedback [25]—using a lab study with DHH users, identifying the promising low-resource deep learning models and smartwatch-based architectures (*e.g.,* watch+phone) for sound classification.

Outside of HCI, wearable vibrotactile approaches have been studied, often for sensory substitution of auditory information [18,29,140,142]. For example, Yeung *et al.* [140] transformed pitch information to vibro-patterns via a 16-channel tactile forearm display. However, obtrusive form factors (*e.g.,* waist-mounted [18], neck-worn [29]) made many of these devices impractical for everyday use. Some early wrist-worn vibrotactile sound aids showed promise (*e.g.,* Tactaid7 [29], TAM [123]), but frequent vibrational feedback had a high attentional cost, especially in noisy environments [102]. Researchers have also tried methods to completely substitute hearing with tactile sensation (*e.g.*, [29]), but with little promise.

2.4 Speech Awareness Technologies

My work is also informed by prior work in speech-to-text systems that provide spoken information access to DHH people using either trained humans [146], automatic speech recognition (ASR) engines, or a combination of both (*e.g.,* [129]). Captions by trained human transcriptionists are highly accurate, but transcriptionists are expensive and require prior scheduling [69]. In comparison, ASR is cost-effective but is error prone, hence recent work has looked at real-time editing of ASR [40] or foregoing ASR altogether by using crowdsourcing [73].

Traditionally, captions from a speech-to-text system are displayed on a laptop or a shared large screen [69]. More recently, mobile solutions have emerged that use ASR (*e.g.*, Google's Live Transcribe [145]) or human-mediated approaches (*e.g.,* [79,134]) to transcribe speech. Though portable, mobile phones require the users to turn their gaze away from the speaker or environment to use the captions. To reduce this visual split, my HMD captioning work examines showing captions directly in the users' gaze area using augmented reality-based HMD displays [50].

As an alternative to captions, Jones *et al.* [59] and Miller *et al.* [90] explored displaying a sign language interpreter via an HMD. In both cases, at least some participants found value in having the interpreter always in their view. However, while in [90] participants found it easier to follow a lecture with an HMD, in [59], a majority found it overwhelming to focus on both interpreter and other study tasks (*e.g.,* watching a movie).

2.5 Sound Recognition Research

My work relies on sound recognition algorithms. Early efforts in classifying sounds relied on hand-crafted features such as zero-crossing rate, frame power, and pitch [101,113,114]. Though these approaches performed reasonably well on clean sound files with a small number of classes, these features fail to account for acoustic variations in the field (*e.g.,* background noise) [75]. More recently, machine learning based classification has shown promise for specific field tasks such as gunshot detection [27] or intruder alert systems [6]. Specifically for DHH users, Bragg *et al.* [13] explored a preliminary GMM-based sound detection algorithm to classify two sounds (alarm clock, door knock) in an office setting. For more broad use cases, deep learning-based solutions have been investigated [72,119]. For example, Sicong *et al.* [119] explored a lightweight CNN-based architecture on smartphones to classify nine sounds preferred by DHH users (*e.g.,* fire alarm, doorbell) in a school setting, and Laput *et al.* [72] used convolutional neural networks to classify sounds in the homes. I closely followed the later approach for HomeSound [54], while adapting for sounds preferred by DHH people in the home. Additionally, as part of the SoundWatch work [55], I trained models for portable devices (phone, watch) using a similar approach and performed evaluations in varying contexts (home, work, outdoors).

While useful, the above approaches rely on generic models pre-trained on large sound corpora which do not support end-user personalization— such as training on new sound categories (*e.g.,* a new custom home appliance) or a the sound (*e.g.,* my child's voice or pet's dog bark)—a highly desired feature by DHH users to support their diversity of needs in different context [13,25]. To support in-situ personalization, relevant machine learning approaches include transfer learning [128], a supervised training method that uses limited training examples to fine-tune a model previously trained on large datasets from a different domain (*e.g.,* image classification). Likewise, co-training approaches [95,131] use a small number of examples and a large unlabeled set to create a model with better classification performance. In my ProtoSound work, I investigate meta learning [132], a learning approach that allows models to recognize previously unseen classes or adapt to new environments with very few labelled training instances. This approach has recently been explored in many domains, including computer vision [26,107], acoustic event detection [116,133], and natural language processing [98,141].

Most similar to my ProtoSound work is ListenLearner [136], which provides a platform for learning new classes through one-shot user labelling. ListenLearner starts with a pre-defined set of classes that it uses to learn representations of new sounds by recording a large number of samples and prompting the user for labelling (*e.g.*, "what sound was that?"). However, this semi-supervised approach requires longitudinal deployments for recording many samples, while our approach allows for quicker adaptation to new environments through fewer training examples. Furthermore, by prompting users for feedback at unspecified times, ListenLearner assumes that users have domain knowledge (i.e., they can listen to and identify a recently occurring sound)—an assumption that may not hold for DHH users. Our intentional recording approach may enable users to leverage visual and contextual cues for recording (*e.g.*, by seeing that a faucet is turned on). Finally, unlike ProtoSound, ListenLearner does not support customizing existing classes (*e.g.*, my dog vs. a generic dog). Acoustic distribution of sound classes may vary widely across acoustic contexts [77] and using existing class representations may not generalize well.

Chapter 3: Investigating Sound Awareness in the Home

The home is filled with a rich diversity of sounds that convey information about the home environment and the occupants within it—from mundane beeps and whirs to children's shouts and dog barks. For d/Deaf and hard of hearing (DHH) people, designing and adapting a home space can mean arranging furnishings, mirrors or lighting to allow for clear sightlines and visual awareness [147], and installing visual and vibrational options for urgent information that is typically conveyed to hearing people via sound (*e.g.*, alarm clock, doorbell, telephone ring).

At the same time, many DHH individuals are interested in having greater access to sound awareness in the home [13,80]. Matthews *et al.* [80] interviewed 18 DHH participants in two studies about sound awareness needs at home, at work, and while mobile, finding that at home, participants were most interested in emergency alarms, people shouting, and appliance sounds. Building on that work, Bragg *et al.* [13] conducted a larger online survey with 87 DHH participants, confirming that emergency alarms, appliance information, and door knocks/doorbell were among the most desired sounds. Further, about half of both d/Deaf and hard of hearing participants in the latter study reported missing a sound of interest daily.

Building on the above work, we conducted four studies to examine how DHH people think about and relate to sounds in the home, solicit feedback and reactions to sound awareness systems, and explore concerns that are relevant to sounds in the home context, such as privacy and activity tracking. In contrast to previous work [13,80], our studies offer more depth on sound awareness needs and preferences in the home, introduce, deploy, and study prototypes that are designed specifically for the home—as opposed to an office environment [13,80] or for mobile or wearable devices (*e.g.*, [88,89]), and evaluate these prototypes in the home.

Study 1 consisted of semi-structured interviews with 12 DHH participants to explore perceptions of and experiences with sound in the home, to learn about sound-related adaptations participants may have made to their home, and to solicit feedback on initial sound awareness mockups. Echoing previous findings [13,80],

we found similar preferences for sounds of interest, sound information to show (*e.g.*, location, loudness), and the devices to display this information (*e.g.*, a wall-mounted device). We also identified the need to show different information for each sound type, the need to select sounds based on the user's location in the home, and themes for further exploration such as privacy, cognitive overload, and mitigating uncertainty of sound classification.

Informed by Study 1 and past work [80], we designed three initial sound awareness prototypes (Figure 3.1) and conducted a Wizard-of-Oz study with 10 DHH participants (Study 2). The study elicited reactions to the prototypes using short real-time demos and thematic scenarios designed around issues of privacy, cognitive overload, and activity tracking. We uncovered several design suggestions to make the sound awareness technology better support domestic processes of DHH people, including contextual factors to control the information shown (*e.g.*, amount of activity in the home), ways to increase the actionability of the system (*e.g.*, showing both start and end times for continuous sounds such as water running) and different visual features to accommodate inaccuracy (*e.g.*, showing source location).

Figure 3.1: Two of the three sound awareness prototypes we used for Wizard-of-Oz evaluations: (a) list and (b) floorplan. Actions performed: (a) liquid pouring, (b) a clap.

Building on the above two formative studies, we then designed and performed two field evaluations (Study 3 and 4) of a home-based sound awareness system called HomeSound—to our knowledge, the first such evaluation (Figure 3.2). Similar to other display-based IoT devices like the Echo Show or Nest Hub, HomeSound consists of a microphone and display, and multiple devices are installed per home; however, device interactions are controlled by touch rather than the user's voice. We iteratively built and evaluated two versions starting with simple but accurate visualizations of sound feedback (Prototype 1) before adding more complex features (Prototype 2). We deployed each version for three weeks in homes with DHH occupants and conducted pre/post-interviews and weekly online surveys.

Prototype 1 was composed of 3-5 interconnected tablet-based displays encased in laser-cut wood frames that sensed and visualized sound characteristics such as loudness and pitch. Upon deployment in four homes, we found an increase in the self- and home-awareness of the participants, who used context (*e.g.*, location, visual cues) to identify sounds from the display visualizations. However, similar to our previous WoZ-based lab study [53], participants expressed the need to incorporate automatic sound identification and alerts. In terms of privacy, the house occupants accepted the always-on monitoring, but some guests voiced concerns.

Informed from these findings, we extended Prototype 1 by adding a sound classifier for 19 common house sounds (*e.g.*, alarms, kitchen appliances) and a smartwatch app for providing alerts about sounds. We deployed this Prototype 2 in four homes (two new, two repeat). Results show a further increase in participants' home awareness. However, misclassification of sounds and frequent smartwatch vibration alerts were not well received.

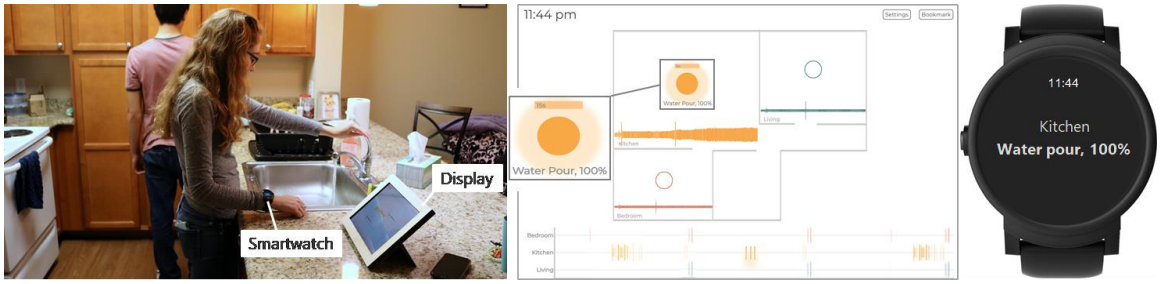


Figure 3.2: HomeSound is an in-home IoT prototype device aimed at improving sound awareness for people who are d/Deaf or hard of hearing. In the image above, an occupant turns off the water faucet after being alerted to the "*water pour*" sound in HomeSound Prototype 2. Screenshots of the IoT display and accompanying smartwatch are also shown. See video.

In summary, the primary contribution of this chapter includes: (1) qualitative insights on DHH people's needs and preferences related to sound awareness in the home, (2) two iterative prototypes of the first home-based sound awareness system for DHH occupants, (3) insights from two three-week field deployments in six unique homes, and (4) design recommendations for future in-home sound awareness technology. Our overarching aim is to help design future IoT devices like the Echo Show or Nest Hub while accounting for the needs and desires of DHH users.

3.1 Study 1: Needfinding and Design Probe

To assess the needs and potential for sound awareness systems in the home, we conducted a two-part study with 12 DHH participants: a semi-structured interview followed by design probes. We first outline a design space that informed the study, before describing the method and findings.

3.1.1 Design Space for Home Sound Awareness

We generated six design dimensions for a home-based sound awareness system based on prior work in domestic computing [6,10] and technology for DHH users [2,14]:

1. **Form factor:** What device is used to convey sound information (*e.g.*, smartphone, wall-mounted display)?
2. **Output modality:** Via what sensory mode does the user receive information (*e.g.*, visual, vibrations)?
3. **Display elements:** What information about sound is conveyed (*e.g.*, sound type, location, length)?
4. **Sound type specificity:** How precisely is the sound type conveyed, from the very specific (*e.g.*, "*fan on low mode*") to moderately specific (*e.g.*, "*fan running*") to more general (*e.g.*, "*whirring sound*")?
5. **Sound location specificity:** How precisely is the sound location conveyed, from very specific (*e.g.*, "*upstairs bathroom sink*") to more general (*e.g.*, "*upstairs*").
6. **Confidence level:** How to convey level of certainty of sound classification (*e.g.*, percentage accuracy)?

3.1.2 Method

Participants

We recruited 12 DHH participants (seven females and five males) through email, social media, and snowball sampling. The number of participants was determined based on reaching thematic saturation (see Data Analysis). Participants were on average 36.8 years old (SD=16.3, range=21–67). Eight participants reported onset of hearing loss as congenital, two reported 1 year, one reported 2.5 years, and one reported 6 years.

Table 3.1: Participants in Study 1 (P1, etc.) and Study 2 (R1, etc.). Counts for total occupants (#Occ.) and for DHH occupants (#DHH) include the participant.

ID	Age	Gender	Identity	Hearing Loss	#Rooms	#Occ.	#DHH
P1	24	F	HH	Profound	10+	7	1
P2, R3	60	M	deaf-blind	Severe	1-3	1	1
P3, R4	55	M	Deaf	Severe	10+	2	1
P4, R1	54	F	Deaf	Profound	4-6	1	1
P5, R7	67	M	Deaf	Severe	7-9	2	1
P6	25	F	deaf	Profound	1-3	1	3
P7	21	M	Deaf	Profound	4-6	1	4
P8	25	F	deaf	Profound	1-3	1	2
P9	25	M	Deaf	Severe	4-6	1	2
P10, R9	33	F	Deaf	Profound	4-6	1	1
P11	32	F	Deaf	Profound	7-9	1	3
P12, R6	21	F	HH	Mild	10+	4	2
R2	45	M	deaf	Severe	1-3	1	1
R5	54	F	HH	Severe	7-9	3	1
R8	21	M	HH	Moderate	4-6	1	1
R10	21	F	HH	Profound	1-3	1	1

Nine participants used digital hearing aids and one used cochlear implants. As our study focuses on the home, we also asked about living arrangements: number of rooms, number of total occupants, and number of DHH occupants.

These details and others, including cultural identities and reported hearing loss level, are shown in Table 3.1. Note that no household had children. Finally, P2, who identified as deaf-blind, reported 20:20 acuity with glasses and peripheral vision less than 20 degrees. He was able to fully engage with our protocol and reported no problems in seeing our visual designs.

Procedure

The study procedure, conducted by the hard of hearing first author, took about 50 minutes. In addition to verbally questioning participants, the discussions were supplemented visually through illustrations, examples, and questions on an iPad. A real-time transcriptionist attended all study sessions, and participants

were given the option of having a sign language interpreter; six participants opted for this accommodation. The study began with a demographic and background questionnaire, followed by a two-part protocol:

Part 1: semi-structured interview. We asked about needs, challenges, and current strategies to access or mitigate the need to know about sounds in the home, as well as ideas for new technologies to address these challenges.

Part 2: design probe. For each dimension of the design space, we provided a brief textual description that included examples of design options (*e.g.*, different display elements or levels of sound type specificity). For the form factor and output modality dimensions, we also presented in random order the five low-fidelity illustrations shown in Figure 3.3A to 3.3E, along with the haptic feedback example shown in Figure 3.3F. Participants chose one or more examples as their preference for a given dimension, and could describe and/or draw a new possibility. We asked for rationale on the choices as well as follow-up questions (*e.g.*, "Would your choice change for different sounds?"). All dimensions were presented in the order listed in the design space above; however, toward the end, we discussed form factors a second time to see if preference had evolved. Finally, because sound sensing is inherently uncertain, we asked whether and how participants would want to see the system's confidence in the sensed sound.

Data Analysis

We applied an iterative thematic coding approach [3] to the session transcripts. Upon receipt of the first six transcripts, one researcher randomly selected two transcripts and developed an initial codebook for each interview section (*i.e.*, challenges, current strategies, future technology ideas, and the six design space dimensions) and identified a small set of emergent codes that applied holistically across all questions (*e.g.*, *privacy*, *information overload*). The researcher then coded the remaining four transcripts, updating the codebook as necessary. After these first six, transcripts were coded one by one until we reached thematic saturation, which occurred at 12 participants. The researcher then performed another pass on all transcripts. The final codebook contained 7-9 codes for each interview section. Finally, two other researchers split the set of transcripts to review all codes, agreeing with the first researcher on 96.5% of the code assignments. The three researchers resolved disagreements through discussion.

3.1.3 Findings

We discuss sounds of interest, existing adaptations and preferences for in-home sound awareness technology.

Part 1: Formative Interview

Sounds of interest: Similar to past work [2,8,13], the sounds of greatest interest were alarms and alerts (*N*=12), appliance timers (10), presence of other people and animals (9), and voices directed at you (7). In contrast, most participants (10) did not want to be aware of continuous background or mechanical noises unless the sound indicates a mechanical problem (*e.g.,* a water leak) or emergency (*e.g.,* a siren) (also identified in [2]). Finally, although participants were interested in sounds indicating presence of people, five participants did not want repeated notifications of that activity, such as creaks, furniture movement, and walking back and forth.

Existing adaptations and challenges: To identify sounds in the home, participants used traditional approaches such as asking for help from other people (5), moving around the house to find the source of perceptible sounds they could not identify (3), and using dogs as guides (3). Several participants also used visual or vibrational alternatives to what are typically auditory devices: doorbells that flash (7) or vibrate the bed (2), a vibratory alarm clock (6), and a wall-mounted light to display the ambient sound level (3).

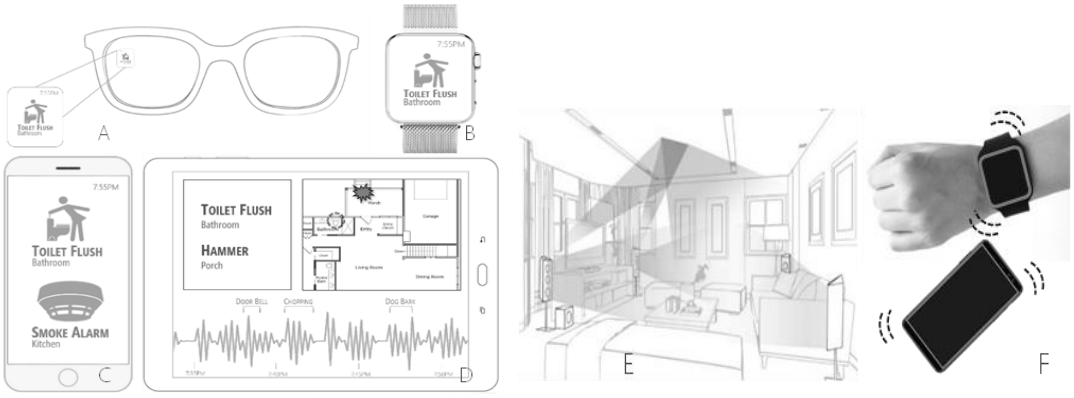


Figure 3.3: Mockups shown to Study 1 participants to guide them in considering form factors (A-E) and output modality (visual A-E and vibration F) dimensions. The form factors are: A: HMD, B: smartwatch, C: smartphone, D: wall-mounted, E: ambient display. The ambient display shines light to illuminate sound sources within a room.

When asked about sounds for which they do not have sufficient adaptations, participants mentioned voices (8) and activity sounds (7). In addition, some participants did not have techniques to deal with mechanical

sounds (5), outdoor sounds (5) or animal sounds (4), but these sound types, as mentioned by four participants, were *"more of nice to haves instead of need to knows"* (P4). Notably, six of the seven participants who used one of the standalone sound awareness devices described earlier (*e.g.*, flashing doorbell) deemed the devices to be insufficient. For example, four participants said they do not like flashing doorbells because the light can only be seen from the room in which it is installed, or the system can trigger falsely. An additional three participants emphasized the hassle of installing multiple devices, such as:

> *"I have [a] flashing doorbell... But, one day I was sleeping and somebody came at night [and] rang the doorbell, and I couldn't see the light. So, I had to get a bed shaker [for the doorbell…] How many devices should [I] keep?" (P11)*

Ideas for future technology: When asked to envision a future technology that provides sound awareness in the home, participants' descriptions most commonly mentioned smartphones (9) followed by wall-mounted or tabletop displays (5). Other ideas included smartwatches, head-mounted displays, subtle notifications on everyday devices (TV, computer), and a ring-mounted device because, *"the phone doesn't help as overnight it is charging in kitchen, but I never take off my ring"* (P5).

3.1.4 Part 2: Design Probe

Form factors: Participants discussed the perceived advantages and disadvantages of the five form factors introduced in the design probe (Figure 3.3). Only three participants liked the HMD, which, among other criticisms, was seen as visually intrusive. We summarize responses to the other form factors in decreasing order of preference.

Smartphone: All participants liked the smartphone because it is portable (6), a device they already own (6), and close at hand (5). Five participants also valued being able to remotely monitor sounds in the home while they are away.

Wall-mounted: Nine participants liked the wall-mounted display, citing benefits of size (compared to smartwatches and smartphones) (6), and static placement in the home (4). However, five participants noted that a downside of the wall display is that it needs to be within their line of sight.

Smartwatch: Eight participants liked the smartwatch, primarily because it is always situated on the wrist (and could be useful for notifications of urgent sounds), and because of its portability. At the same time, the smartwatch was less preferred than the smartphone because most participants do not already wear one (10) and it has a smaller screen (3).

Ambient: Six participants liked the ambient display, where objects making sounds within a room would be visually lit up. Advantages cited by these participants included that the ambient display would offer good visibility and would convey specific sound sources well. For example,

> *"[Unlike a smartphone], you don't have to try and figure out where the sound is coming from. You can tell immediately. Like if you are here in the room, you will see the light from the [source] and you will know where the sound is coming from and whether you should pay attention to it." (P1)*

However, other participants (4) worried about cost and installation effort for the ambient display as well as visual clutter, and so only wanted to use this form factor for some specific sounds (*e.g.*, alarms and doorbells).

Output modality: All participants wanted visual information for all types of sounds, but 11 of the 12 also desired vibration for when immediate attention is needed or while sleeping. Six participants felt that vibration to convey the presence of sound paired with a visual display for more information would be useful. We also asked about olfactory feedback, which five participants were receptive to, mainly for emergencies and alerts (5). For example, P12 said: *"If I'm in a different room and there's a smoke alarm sounding, […] a powerful blast of scent could grab my attention."* However, four participants did not want to use smell.

Display elements: We asked about six potential display elements: *sound type*, *location*, *temporal history*, *length of occurrence*, *physical characteristics* (*e.g.*, loudness), and *importance*. All participants wanted *sound type* and *location*, reflecting past work [14], while at least eight participants wanted each of the other elements. However, eight participants also felt that extra details would be unnecessary for sounds occurring in the same room as the user. When asked about the utility of these elements for different types of sounds, sound type and location were seen as sufficient by many participants (7) for sounds requiring immediate attention, namely, alarms, alerts, and voices. The additional information could be useful, however, to provide context for activity sounds (7), such as the length of a sound being related to urgency:

"Like for footsteps: how long are they heard? If my housemate is pacing back and forth, it could be something wrong. That could provide me an opportunity to check if everything is alright." (P3)

Sound type specificity: No clear pattern emerged, indicating a need for a more grounded evaluation, as supported by P11: *"This is one of those things where I feel I would need to try a system on my own and try filtering it and testing it."* We revisit this design dimension in Study 2.

Sound location specificity: Most participants (8) wanted locations to be at least moderately specific (*i.e.*, to display the room in which the sound occurred), for example: *"I have a big house [...] if somebody is knocking, which door, front door or back…"* (P3)

The remaining participants felt that the need for location specificity would depend on the importance of sound: all wanted a more specific location for more important sounds.

Confidence level: Participants were evenly split on whether to display the system's confidence level of the sensed sound. For example, P8 wanted the confidence level: *"because I am curious [on] how well the technology is going to work. And how I can improve [the system] to detect better in future,"* whereas P1 did not need the confidence level because she: *"would only use the technology if it's accurate [in detection]."* As with sound type specificity, a more grounded evaluation may be useful.

Other considerations: Several emergent themes pointed to issues unique to a sound awareness device in the home. For example, P7, who lived with three other people, felt that his preferences for sound awareness *"would be different if I lived in a house by myself.*" Privacy was also mentioned by six participants, which included interpersonal privacy concerns. P1, who lived with six hearing adults, said: *"I don't want wall mounted in other people's rooms [in the house]. Only maybe in my bedroom and bathroom because I only need to know this information."* Finally, comments about information overload, a concern for eight participants, included that continuous, familiar sounds in the home may not need to be displayed:

"If it's going to remind me every five minutes that tools are making noise in the workshop and that there's a motor sound in the bedroom, that could be very annoying" (P6).

Study 1 Summary

Participants appreciated the idea of sound awareness in the home, and suggested showing alarm and alerts, presence of other people and animals, activity sounds and voices directed at them. Smartphone and wall displays, and a combination of visual and vibration modalities were the most preferred. Participants suggested customizing what sound information to show (*e.g.*, location, length of occurrence) based on the type of sound. Finally, themes of privacy, multiple occupants, and information overload emerged, which we explore further in Study 2.

3.2 Study 2: Wizard of Oz Evaluation of Initial Prototypes

To gain further insight into in-home sound awareness technology, we designed and performed a Wizard-of-Oz evaluation of three sound awareness prototypes in a home-like space (see Figure 3.4b for layout). These prototypes enabled further investigation into themes that emerged in Study 1 but were difficult to study through interviews and static design probes.

3.2.1 Wizard of Oz Prototypes

Informed by Study 1 and past work [14], we created three web-based prototypes that employ different approaches to displaying sounds and that included elements meant to elicit discussion on themes such as privacy and location. The prototypes, also shown in Figure 3.4, were:

List: This prototype displays sound activity using a scrolling list. New sounds appear at the bottom of the list with the time of occurrence and the room location. A search bar at the top allows the user to search for specific sounds, and thus access history.

Floorplan: This prototype shows the spatial layout of the house. Sensed sounds appear as fixed-sized blue bubbles for three seconds within the room where the sound occurred. This instantiation demonstrates moderate location specificity, at the room level, but also allowed us to discuss other levels of specificity (house, specific object in room). Sounds from outside of the house appeared to the right of the spatial layout.

Waveform: This prototype includes a continuous scrolling waveform of sound level (loudness) based on a single microphone as input, with text describing the recognized sound and location above the waveform. This

design allowed us to investigate the perceived utility of physical sound information, as well as the possibility of only showing a waveform, which is relatively easy to implement, rather than more advanced sound identification. While perhaps less approachable than the other designs, as users gain experience they may be able to extract information from the waveform's depiction of physical sound.

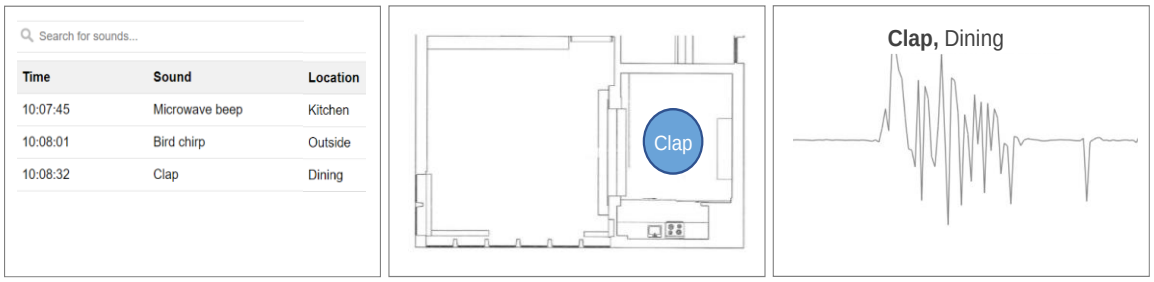


Figure 3.4: The three prototypes used for Study 2 Wizard-of-Oz evaluation: list, floorplan, and waveform

The Wizard of Oz setup consisted of a Microsoft Surface Pro 3 tablet, which displayed the prototypes in a web browser, and an HTML-based wizard interface that ran on a separate laptop and communicated with the tablet over wifi via a web socket connection. The wizard interface included location buttons the four rooms at the study setting (dining, kitchen, bathroom, lounge, outside) and a list of pre-set sounds that were likely to occur during the study, along with a textbox to enter other sounds. Additionally, for the waveform prototype, we connected a conference microphone to the tablet and used Mozilla's Audio API to generate a real-time waveform.

3.2.2 Method

Participants

We recruited 10 DHH participants (five females; five males) through email lists and snowball sampling (Table 3.1). Because this study extended Study 1, we did not exclude repeat participants. As a result, six Study 1 participants also volunteered for Study 2. Participants were on average 44.1 years old (*SD*=15.7, *range* 21–67). Seven reported onset of hearing loss as congenital, and the other three reported 1, 6 and 15 years. All participants used digital hearing aids. Most (7) reported living in a home or an apartment with at least 4-6 rooms. Four participants lived with at least one adult; one of the occupant's in R5's home was a child.

Procedure

Study sessions were held in a student and faculty lounge on a university campus. This setting was chosen to offer control across participants but also to be home-like: it contained a kitchen, dining room, lounge area (like a living room), bathroom, and windows to the outside. The protocol was designed to take an hour, including a written background questionnaire and three longer activities: initial prototype demos, thematic scenarios, and a semi-structured interview. Similar to Study 1, the protocol was conducted verbally with pre-planned questions also presented on an iPad. A real-time transcriptionist attended all sessions. Participants were additionally offered sign language interpretation if desired, which two participants ultimately requested. Three research team members were present during the sessions: the lead facilitator, a wizard, and an actor. The participant, facilitator, transcriptionist, and interpreter (if present) sat at the dining table, while the wizard sat on a couch in the lounge area behind the participant and the actor moved around the area as needed.

Part 1: initial demos and design ideation (10 mins): The three prototypes were demonstrated in random order. These Wizard of Oz demonstrations were designed to briefly introduce participants to the features of each prototype and to give a sense for how the prototype worked for a small set of everyday domestic tasks. The actor performed three sets of everyday actions: (1) Starts the microwave then does dishes in the kitchen [microwave starts, beeps after 10 seconds, water running, dishes clinking]. (2) Knocks at the door [door knocking], which is opened by the facilitator [door open]; after greetings [speech (greetings), door close], sits next to the participant. (3) Makes coffee [machine starts, stops, liquid pouring, object placed], while bird chips outside [bird chip]. For each participant, these sets were randomly paired with the three prototypes. Participants were also encouraged to make their own sounds (*e.g.*, clap, table tap). After the demos, participants explained their reactions to each prototype, suggested improvements (if any), rated the usability of the prototypes, and were invited to sketch new design ideas on blank templates of the prototypes.

Part 2: thematic scenarios (20 mins): Following the initial demos, three new scenarios were presented, one for each prototype. These scenarios were designed to explore themes that could manifest in unique ways in the home setting: privacy, activity tracking, information overload, use of spatial layout, and sounds of interest. Each scenario was first described as follows (without the bracketed sounds), then for the bathroom and movie scenario, the actor played out the scenario as the prototype displayed the sounds:

Bathroom Scenario: You are in your dining room with a fa mily member. You are reading a book. The family member gets up and goes to the bathroom [footsteps, door open/close] while you continue reading. [bathroom sounds (*e.g.*, toilet flush, water running)] The person comes back but forgot to fully close the sink faucet [water dripping].

Babysitter Scenario: You come home from work at 5PM. While you were gone, you left your kids with a babysitter. You're interested in knowing the activities in your house that occurred since your baby woke up from its nap at 4PM. You go and check the history of sounds. [Example sounds include baby crying, baby laughing, toys rattling, liquid pouring, singing].

Movie Scenario Imagine you are in your dining room working on your laptop [keyboard typing]. It is a hot summer evening, so you have your AC on [AC hum]. Your roommate is in the lounge watching television while on the phone with their partner [Sounds from TV (*e.g.*, music, children playing, street sound) as well as phone ringing, roommates' speech, furniture sounds].

We paired each scenario with a default prototype that would better generate reactions to the specific themes we wanted to explore in that scenario. For example, pairing the babysitter scenario with the list prototype (that shows history) allowed us to study the theme of "activity tracking". Similarly, the bathroom scenario was paired by default with the floorplan prototype, and movie with the waveform prototype. However, all three prototypes were open in the browser for each scenario and we encouraged participants to explore them. For the bathroom scenario, the actor waited outside the study location and the wizard produced sounds as if the actor were using the bathroom. For the babysitter scenario, 18 pre-selected sounds were shown (examples above). We elicited participants' initial reactions after each scenario.

Part 3: semi-structured interview (30 mins): Finally, we asked semi-structured questions about the participants' experience with the prototypes, focusing on information overload, privacy, and activity tracking. We also showed three additional simple mockups for conveying system uncertainty in sound sensing, to further examine this theme from Study 1: showing only the *location* of a sound within a room but not the *type* of sound (from [14]), showing a list of possible sounds with accuracy, and showing only a waveform of sound loudness.

Data Analysis

We analyzed the professional transcripts using a process similar to Study 1. One researcher selected two transcripts to develop an initial codebook, then iteratively applied the codebook to all transcripts (while refining the codebook). The final codebook contained 6-8 codes for each of the five sections—three sections of the transcripts and two additional themes (*personalization*, *installation*) that emerged across the entire transcripts. A second researcher then conducted a peer review of all coded data, agreeing in 96.0% of cases with the initial researcher. Disagreements were resolved through consensus.

3.2.3 Findings

All participants reacted positively to the idea of sound awareness feedback in the home. For example, R4 on seeing a 'bird chirp' visualization said: *"I can only hear birds in the forest alone when nobody is around and I turn my hearing aids all the way up. Cool!"* Similarly, R5 said:

> *"I get real anxiety when my husband is traveling for work. Our bedroom is on the first floor and [our] daughter's room is downstairs. I wouldn't hear if she needs my attention. To have something like this in my home would be amazing." (R5)*

However, participants also highlighted challenges with using a sound awareness system in the home. We discuss reactions to our prototypes and insights on the high-level usability themes that arose (*e.g.*, privacy, trust).

Sound Awareness Prototypes

When discussing their preferred designs, nine participants selected the *floorplan* because it intuitively visualized the position and type of sounds. For example, R2 stated: *"This is the best option, because you can see where the sound is coming from and you know exactly where to look."* Three participants valued its glanceability, in that they could notice the sound bubbles without paying active attention.

However, because this display always visualized sounds in the center of a room, all nine participants suggested finer-grained localization. *"I want to see door knock on top of the door, not where it shows now [in the center of the lounge]"* (R3). Other suggestions included using size or color to indicate loudness or

pitch (4), or rings to indicate loudness, pitch (3) (see Figure 3.5A), and showing the time of occurrence (3). For example:

> *"Some deaf people may not know the difference between banging and tapping [on the door]. Having the blue [disk] change to red [to show loudness] would [solve] that." (R7)*

The second most preferred design was the *list* prototype, which was R9's top choice and the remaining nine participants' second favorite. Participants liked that they could easily see a sound's exact timestamp (6) as well as a recorded history (5). Unlike the floorplan prototype where visualizations are transient, the list prototype records and displays sounds chronologically: *"I am going to be busy doing other things and I don't have to check this out every 3 seconds"* (R7). However, while the floorplan view was considered to be glanceable, five participants felt just the opposite about the list display, particularly when *"there are many sounds in the list"* (R9).

Finally, no participant selected the *waveform* prototype, finding that it required active visual attention (7), lacked information (4), and was hard to understand (3). However, seven participants saw the *"benefit of using [waveform] for loudness information"* (R2) and suggested combining it with other prototypes—by displaying a waveform in another column of the list prototype (Figure 3.5B), at the bottom of the floorplan prototype, or in the bubbles of the floorplan prototype. Similarly, six participants suggested a hybrid of the floorplan and list. Table 3.2 summarizes the questionnaire responses on the prototypes, which confirms that the waveform was not as well-received as the others.

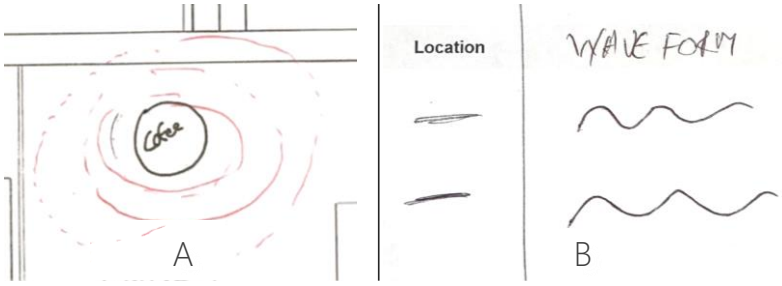


Figure 3.5: Design improvements sketched by participants. (A) R8 sketched rings around the sound bubbles of floorplan prototype to indicate loudness. (B) R10 added a column in the list design to show the sound waveform.

Table 3.2: Mean (and SD) of participant responses to 5-point Likert-scale questions (5 is best), showing that the floorplan and list prototypes were preferred to the waveform prototype.

	Floorplan	List	Waveform
Understandability	5.0 (0.0)	4.9 (0.3)	3.9 (0.9)
Addresses needs	4.8 (0.4)	4.3 (0.6)	2.4 (1.3)
Likely to use in own home	4.5 (0.9)	4.2 (1.0)	2.1 (1.2)

Home-based Sound Awareness Themes

We now describe high-level themes that arose related to usability of a future in-home sound awareness system.

Actionability: All participants emphasized that a sound awareness system could help them perform desired tasks. Several suggestions for ensuring actionability arose, including: showing physical characteristics for sounds that require an action (8) (*e.g., "[distinguish] loud banging from a soft door knock using loudness"*, R7), using vibrations on smartwatch or phone to notify about sounds (7), identifying that someone is calling for your attention as opposed to general conversation (7), and showing additional information for voices (*e.g.,* tone) to identify if they demand attention (6). For example, R5 emphasized critical differences in sounds from her daughter's room:

> *"If my 9-year old daughter is sitting down in her room and talking normally, I am not concerned. If she's sobbing because she's in pain, then I need to know that."*

For continuous sounds, because we only showed the beginning time, participants also wanted to know the end time to avoid needing to close the sound source (5):

> *"say a water flow… if it had a beginning and no end then that would be a problem. Perhaps have the circle stay on [in floorplan prototype] till the water stops flowing…" (R7)*

Finally, three participants mentioned the importance of distinguishing real-life sounds from digital media like movies or music, as captured by R7: *"I saw a 'smash' [on display]. I better go check if some-thing fell on ground but I don't know it's from TV."*

Trust and confidence: Without trust, a sound awareness system is purposeless. We asked participants about three types of sensing errors that could arise: misattribution, false negatives, and false positives. Misattribution, that is, misrecognizing one sound for another, was seen as the most problematic, with most participants (7) feeling that these errors would undermine their trust in the system. However, false negatives, that is, not showing a sound that *had* occurred, would not pose a significant problem unless the sound was safety related (*e.g.*, fire alarm) (7). For example,

> *"Whatever is more important and signifies a dangerous moment or something that is really life driven [should not be omitted]. I was at home alone and the carbon monoxide [alarm] was beeping for one week. My neighbor told me. That's when I corrected it. Those are important kinds of sounds I'd like to know. I don't mind if it [occasionally] misses out on sounds that are not important, because I've lived all my life like that." (R10)*

Finally, most participants felt that false positives—showing a sound that did not occur—would be tolerable but annoying. For example,

> *"I am fine with it […] But, if it repeatedly shows a sound coming from somewhere, and I keep on checking checking checking, we have to figure out what is going on, you know?" (R9)*

After viewing the three options (location, list, waveform) for how to handle the system's uncertainty in identifying sounds, all participants said they would use the system even if it shows only a sound's *location* but not its *identity* (*e.g.*, bathroom but not toilet flush); this finding reflects past work [14]. Eight participants also felt that the location plus other uncertain information could be useful, such as a high-level category (*e.g.*, *"an alarm-like sound"*, R8) or a list of possible sounds (*e.g.*, *"this sound could be a clock alarm or a microwave beep"*, R4). As with the overall Wizard of Oz prototypes, the waveform was not well-received: eight participants felt that sound characteristics such as loudness would not be useful without location. Finally, only four participants wanted to see the confidence level (as shown in the list mockup) because it could reduce trust in the system and is not meaningful information. For example, R7 said: *"If it said 52% microwave beep or 23% something else […] you are still going to go and check which [sound] it is."*

Finally, four participants wanted the system to provide an option to manually correct mistakes (*e.g.*, to classify an unrecognized sound), for example, if a hearing friend noticed that the system was incorrect (R9).

Privacy: Privacy issues arose related to intimate sounds, activity tracking, and unwanted access to sound history. Toward the first issue, four participants felt uncomforatble about the system showing bathroom sounds. For example, *"I don't want to know if someone is using toilet or what ever they are doing in the bathroom… It's their privacy, you know?" (R8).* Conversely, six participants were more open to these sounds for accessibility reasons. For example, R1 emphasized the need for equal access to information: *"Hearing people can hear all that, right? [So,] it's fair to have equal access to information. I want it all."* There may be a cultural component to these preferences: the four participants who identify as hard of hearing were more concerned, but the five D/deaf participants and one deaf-blind participant were more open to these sounds.

As a second dimension of privacy, four participants noted that a sound awareness system could provide insight into other household members' or guests' activities—which may not be desired by either party. For example, *"people [would] avoid coming to my house because they're being monitored each and every moment…"* (R10). However, most participants (9) saw some value in tracking, for example, *"I want to know what my cleaning lady's been doing. If she is using my computer without permission, then I can [know]"* (R5), and *"The value of this in monitoring a baby is significant"* (R4). Consequently, to overcome privacy issues with activity tracking, some participants suggested including a setting in the system for selective sound recording (7) and letting other people know that their sounds are being recorded (5).

A final concern was household members or guests accessing the participant's past activities, particularly when a sound awareness display is installed in a shared space (6). For example, *"I was thinking if this was installed in the living room, and some [guest] comes in, they can see what I've been doing?"* (R8). Thus, two participants wanted the system to have a password option.

Information overload: A sound awareness system could lead to overload if many sounds occur simultaneously. Indeed, in the movie scenario, where participants viewed sounds while pretending to work on a laptop, they felt overwhelmed (9) and distracted (6). Thus, nine participants wanted the system to filter sounds and gave suggestions for doing so. For example, R7 wanted detailed customization:

"If I can enter [in the system] if [it shows] chair squeaking: okay I don't need to know that in future. Don't show that again… Also, after 5 days I can go back and select sounds from a list, like I don't need to hear this but I want to continue hearing this." (R7)

As in Study 1, five participants did not want to be aware of background noises unless it indicated a problem, such as the AC system clicking instead of humming (R5). Other suggestions to reduce overload included: limiting the system to mostly actionable sounds such as safety-related sounds (6) and human voices (5), and using colors to filter out sounds based on importance (5) (*e.g., "red is fire alarm, blue is water [running]"*, R3). Finally, four participants suggested using a manual "sensitivity" setting so the system can automatically filter sounds below a decibel level.

Contextualized feedback: Participants mentioned several contextual factors related to the home that affect sound awareness preferences: daily rhythms, domestic activities, and location in the home. In terms of daily rhythms, six participants wanted the system to limit sounds based on the time—for example, at night:

For night time you might want to [show] crying if the kids are in the other room. You wouldn't worry about siren outside or street noise or the air conditioning running. Those are daytime things. […] Now, television. If it is happening between 10 o'clock at night and 5 in the morning [then I need to know]" (R5)

Four suggested adjusting the system based on domestic activities, including the user's current activity (*e.g.,* only display urgent sounds when the user is working, R9) or the amount of activity in the home, such as: *"Restrict to important sounds only when there's a large guest party, [because] I don't want to be distracted at that time"* (R4).

Three participants wanted to adapt the system based on the user's current location. For example, R5 mentioned not wanting notifications for sounds when she was in a position to receive the same information visually (*e.g.,* not needing a doorbell sound when she is in view of the front door). R2 mentioned not necessarily wanting to know about cooking sounds when he is cooking, but when "*I am in bedroom and somebody is cooking, those sounds are important."*

Other design considerations: Interestingly, participants were split on whether they wanted the system to show their own sounds. While five participants did not want to be aware about their own activities (*e.g., "If I am [operating a faucet], I know that I am running water,"* R9), four participants felt they would benefit from knowing if they were making a loud noise or doing something incorrectly:

> *"I have had troubles when I talked loudly or played music loudly for other people. So, if [the system] could tell me the music is up too loud [that would help me respond]" (R5).*

Finally, three participants suggested an option to train the system to identify their custom sounds, which supports work on personalized sound sensing [2]. For example,

> *"If I get a new microwave then the sound is different. I want to tell [the system], here's what my new microwave sounds like, [and] so, it can learn from it." (R1)*

System installation: A future sound awareness system should seamlessly integrate into the home. Five participants emphasized the importance of accommodating existing home adaptations such as light-based technologies and vibrating devices. For example,

> *"The [vibrating] device under my bed could talk to the smartphone and vibrate if there's an emergency." (R3)*
>
> *"I already have light [based tech] at my house. I want a coordinated system that [includes] these devices," (R10)*

Some (5) wanted the system to integrate with other smart home systems. For example, R9 discussed using a sound sensing system alongside her security camera:

> *"I would look at [list prototype] first because I can scan it much more quickly than having to go through the security camera [recordings]. For instance, if I go through this list and see that the child cried for 30 minutes or something, then I can sync up with the video with the time stamp to see what went wrong." (R9)*

Finally, participants were concerned about the cost and effort of the system installation and gave useful suggestions. All participants wanted to standardize the process of obtaining floorplan data to minimize

overhead (*e.g.*, by using templates). Six participants wanted to install the system in only the most-used rooms in the home to lower the cost. Finally, R4 and R2 suggested to show the *"direction of the sound source instead of the specific location" (R4)* if obtaining floorplan becomes difficult.

Study 2 Summary

Participants appreciated the spatial layout and glanceability of the floorplan prototype, though also liked the list prototype; the waveform prototype was not well-received. We uncovered ways to increase the actionability of the system (*e.g.*, showing both start and end times for continuous sounds) and to mitigate uncertainty of sound sensing (*e.g.*, avoid misattribution errors, show location even without identity). Other themes arose, such as privacy related to other occupants or visitors, the need to consider contextual factors (daily rhythms, domestic activities, and location), and a desire for future systems to work with existing home adaptations and smart home technology.

3.3 Study 3: Prototype 1 Field Deployment

Informed by the above two formative studies, we designed and performed a field study of the first working prototype of our in-home sound awareness system: *HomeSound*.

3.3.1 HomeSound Prototype 1

HomeSound is inspired by commercially available display-based domestic IoT devices like the *Echo Show* or *Nest Hub* but designed specifically to provide sound information to DHH users. To create HomeSound, we followed a human-centered iterative design process starting with the construction and evaluation of a simple but accurate sound feedback prototype (Prototype 1) before building and deploying a more complex system (Prototype 2). With Prototype 1, our goal was to examine how DHH users and other home occupants would react to and experience a sound awareness system, which conveyed four sound properties: room-level location, loudness, duration, and pitch.

Prototype 1 consisted of 3-5 interconnected "picture frame" displays (Microsoft Surface tablets encased in a laser-cut wood frame). Each display continuously sensed, processed, and uploaded sound information in real-time, which was further processed by a backend server to produce a single across-home sound feedback visualization. Though the tablets were general purpose computers, the HomeSound displays were intended

to function as IoT devices—no other tablet-based applications or interactions were possible. Below, we describe HomeSound's privacy-preserving sound sensing pipeline, visualizations, and our implementation.

Sound sensing pipeline. For domestic IoT systems, privacy is a key concern [42,53]. While HomeSound relies on a distributed set of live microphones, from the onset, we designed our sensing pipeline to protect user privacy. Each device processes sound locally and only uploads non-reconstructable features. For signal processing, we take a sliding window approach: HomeSound samples the microphone at 44kHz and segments the data into 50ms windows (2200 samples). To extract *loudness* and *frequency*, we compute the average amplitude and maximum frequency in the window (FFT bin size: 20Hz; range: 0-22kHz) and upload the results. For each display, the backend server stores this information in a database and applies a simple *sound event* detection algorithm: when loudness crosses a minimum threshold (46dB), a 'start' event is marked, which then 'ends' when loudness falls below 46db for one second. These thresholds were determined during a one-month pilot in home of the first author, who is hard of hearing.

Visual display. Informed by previous work [53,80], we designed the HomeSound display to be simple, glanceable, and require no direct interaction. The visuals are composed of two primary views: a *floorplan* (top half) and a *history* view (bottom half). In addition, a header bar displays the current time and a bookmark button, which allowed users to mark an event of interest for consideration by the research team (when pressed, the system took a screenshot and opened a form for typed feedback) (Figure 3.6).

The *floorplan* view showed a top-down blueprint of the home, which was overlaid by real-time sound information. For rooms with an installed HomeSound device, a 'pulsing' circle displayed in the room's center depicted real-time ambient sound loudness—the circle's radius was drawn proportionally to sound amplitude. At the top of each circle, we displayed a 'sound event duration bar', which visualized the length (in time) of the currently detected sound event. To enable comparisons across time, two circle outlines were drawn on top of the pulse showing average room loudness for the past 30 mins and 6 hours. In addition, we displayed a 30-second scrolling waveform at the bottom of each room, intended to help users detect visual patterns in sound activity.

For the *history* view, we created a custom time-series visualization, which showed per-room sound activity over the last six hours. Inspired by [148,149], sound events are displayed as rectangular blocks—block width represents duration, height is average loudness, and color opacity is pitch. The six-hour window was selected to enable recent comparisons across time (10 secs = ~1px) while balancing privacy concerns related to longitudinal patterns.

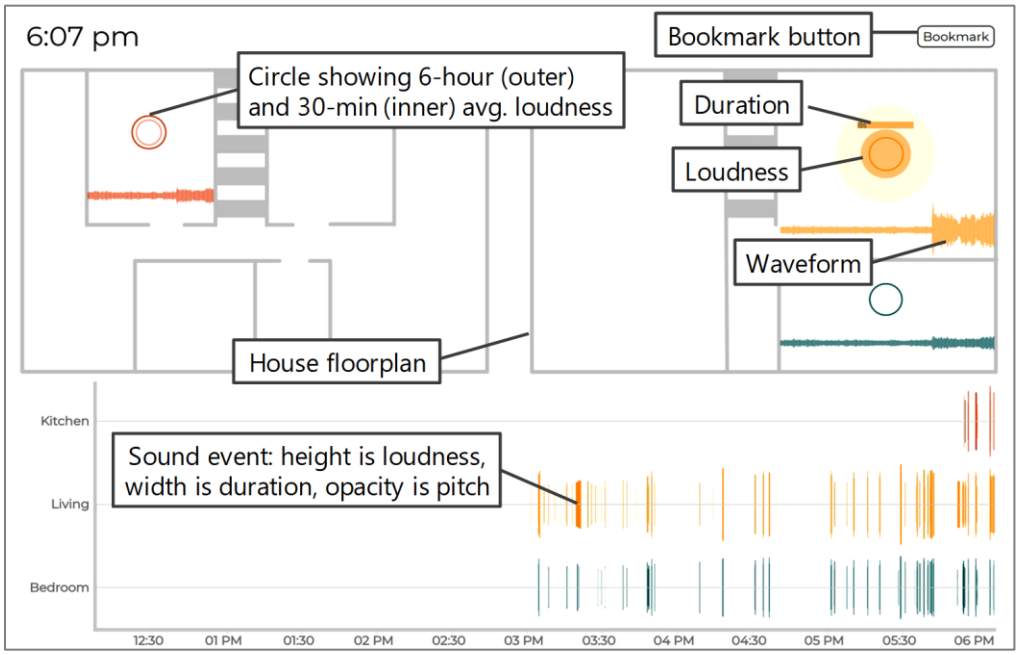


Figure 3.6: HomeSound Prototype 1 interface showing the floorplan view (top half) and history view (bottom half).

Implementation. We implemented HomeSound in Node.js [127] using a client-server architecture composed of three parts: a data client, a web client, and a backend server. For the displays themselves, we used Microsoft Surface Pro 3 tablets (i7 1.4GHz, 4GB RAM) encased in a custom wood frame, which ran both the data and web clients. The data client sensed, processed, and uploaded sound data to the backend server while the web client visualized sound information downloaded from the server in a full-screen Chrome browser. We used pyAudio [150] for sound processing, socket.io for client-server communication, and D3 [12] and CSS for the visualizations. For the backend, we built a Node.js HTTP server on a Windows desktop computer (Intel i7 running Windows 10) using a pm2 process manager [151]. For each home, the server received pitch and loudness data from the data client, computed sound events (loudness, pitch, duration), and

broadcasted this information in real-time to all web clients. Because the sound data (and state information) was stored on the backend, new web clients could be easily launched and supported.

To examine how DHH users react to our prototype, we performed a field study in four homes.

3.3.2 Method

Participants

We recruited DHH participants using email, social media, and snowball sampling. As our study focused on the home, we also recruited hearing household members of the DHH individuals. Six DHH and one hearing individual agreed to participate, and they were on average 62.4 years old (*SD*=12.8, *range*=43-79)—see Table 3.3. Of the six DHH participants, four reported congenital hearing loss, H1P1 reported onset at 3 years old and H3P1 at 4. Two participants used digital hearing aids and one used cochlear implants.

Table 3.3: Homes for Study 3 (H1, H2, H3, H4) and Study 4 (H1, H2, H5, H6) with participant characteristics. Counts for occupants in the home (#Occ.) and for DHH occupants specifically (#DHH) include the participant.

Home	ID	Age	Gender	Identity	Hearing Loss	#Rooms	#Occ.	#DHH
H1	P1	67	M	HoH	Severe	4-6	2	1
	P2	77	M	Hearing	N/A			
H2	P1	79	M	Deaf	Profound	7-9	2	2
	P2	60	F	Deaf	Profound			
H3	P1	56	M	HoH	Profound	10+	4	2
	P2	55	F	HoH	Severe			
H4	P1	43	M	Deaf	Profound	4-6	3	1
H5	P1	50	M	Deaf	Profound	7-9	4	2
	P2	49	F	Deaf	Profound			
H6	P1	22	F	HoH	Severe	4-6	3	1
	P2	21	F	Hearing	N/A			

Procedure

The study, conducted by a hard of hearing author, had three parts: an initial interview and system installation, three-week system use, and a post-trial interview. Both pre- and post-interviews were held in the participants' homes and audio recorded. A real-time transcriptionist attended all interviews, and five participants opted to

also have a sign language interpreter. Consent and background forms were emailed in advance; written consent was taken in person.

Part 1: Initial session (1 hour). The initial session began with a 20-minute semi-structured interview (5 questions) with the DHH participants on experiences with sounds in the home, challenges faced, and coping strategies. The hearing participants then joined for a PowerPoint presentation on HomeSound, including the visualization overview and how user privacy is preserved. Afterwards, the participants gave a brief tour of their home and discussed the display placements. The initial number of displays was based on the home size (three for homes of <1000 sq. ft, four for 1000-1500 sq. ft, and five for >1500 sq. ft), but participants could ask to add or remove a display during the study. Though participants could choose any locations for these displays, we provided three suggestions: kitchen, living room and entryway. After the tour, the researcher took 20 minutes to draw the floorplan using online software and uploaded it to the server. The displays were then placed in the desired rooms on a flat surface (*e.g.*, kitchen counter, bedside table) based on visibility and proximity to a power source. The researcher initialized and demoed the system by making some sounds (*e.g.*, clap, speech) in front of each display. Finally, participants were provided a UI reference sheet, and encouraged to give feedback using the bookmark form.

Part 2: Deployment period (3 weeks). During the three-week deployment, participants were instructed to perform their usual daily activities, interacting with HomeSound if desired. We emailed three weekly surveys (5 open-ended questions each) about overall experience, sound awareness, and any positive or negative incidents. If a participant did not complete a survey within 24 hours, a reminder was sent.

Part 3: Post-trial interview (1 hour). At the end of the deployment, we conducted another one-on-one interview with each participant (20 questions for DHH and 10 for hearing participants) on their system usage and experience, sound awareness, concerns, privacy issues and design suggestions. We also asked follow-up questions based on system logs, bookmarks, or survey responses. After the interview, we retrieved the displays.

Data Analysis

We conducted a thematic analysis [14] on the interview transcripts and weekly survey data. One researcher skimmed the transcripts to familiarize themselves with the data, and conferred with the research team to generate an initial codebook. The researcher then iteratively applied codes to all transcripts while refining the codebook. The final codebook contained a 3-level hierarchy (11 level-1 codes, 54 level-2 codes, and 108 level-3 codes), of which the level-1 codes formed the high-level themes. Another researcher used this final codebook to independently code all transcripts. Interrater agreement between the two coders, measured using Krippendorff's alpha [68], was on average 0.66 across all questions (*SD*=0.30, *range*=0.47-1.0); raw agreement was 86.3% (*SD*=11.7, *range*=70.4-100). Though the alpha value borders the acceptable minimum (0.667), the two coders resolved all disagreements through consensus.

3.3.3 Findings

We cover overall usage, sound awareness, display placement, privacy, and design suggestions. Throughout, we refer to the six DHH participants and report quotes from the post-trial interview unless otherwise noted.

Overall usage patterns. On average, each home had 2411.8 total sound events/day (*SD*=689.5, Table 3.4). Participants completed all weekly surveys, created 46 bookmarks, and sent feedback using 9 email threads, and 21 text messages. Complementing this quantitative data, all participants reported viewing the HomeSound displays at least a few times a day, both explicitly (*e.g.,* to review past sounds) and incidentally (*i.e.,* noticing it during other activities). For explicit use, all participants reported checking a nearby display every few hours, and almost all (*N*=5 out of 6) reviewed sound activity when they came home from work. All participants also mentioned noticing sound information while walking around the house or engaged in other activities—particularly activities that generated sound (*e.g.,* cooking, conversation, laundry). Perhaps unsurprisingly, all participants reported decreased usage over time, as is evident from the logged bookmark data:

> *"I looked at it a lot in the first week, but then not so much in the end. I got used to its presence and forgot it was there…" (H4P1)*
>
> *"I felt like it was fun to look at it initially but then I am so used to living without sounds, that I used it less in the end." (H2P1)*

However, emphasizing the utility of the system for some people, H1P1, H3P2 and H4P1 mentioned feeling nostalgic about it after the deployment period ended, such as,

> *"I was waiting for the microwave to beep but was in the BEDROOM. So, I asked [my husband] if he could hear [it] […] And he said: "where's the system!?" (H1P1, text sent post-study)*

Table 3.4: Study 1 homes, with sizes (in sq. ft.), number of displays, the most active room, daily average of sound events, and total bookmarks for each of the three weeks (B1, B2, B3).

Home	Size	Floors	Displays	Busiest	Events	B1	B2	B3
H1	1060	1	4	Den	3383.4	8	4	1
H2	1740	2	5	Dining	2041.7	5	2	0
H3	2700	2	5	Family	1545.3	11	6	3
H4	950	1	3	Kitchen	2676.6	4	2	0

Sound awareness. In terms of sound awareness, four participants reported how HomeSound made them realize that they were previously unaware of many sounds in their home. H3P1, for example, wore a hearing aid but said, *"I knew I was missing certain sounds. [But] I didn't know how much I was missing"* (interview). All participants reported combining information from the HomeSound displays with contextual cues to determine sound activity:

> *"Every time I walked around the house, I saw disks [pulses] on tablets [emanating from] multiple rooms. I realized that my whole wooden home makes a lot of noise" (H3P1, week 1 survey).*
>
> *"The peaks in waveform from kitchen meant that the microwave must have beeped, and my food was ready. […] No one else [was] in the home." (H4P1, week 2 bookmark, see Figure 3.7 below)*

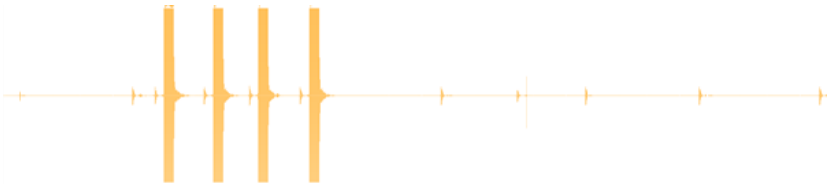


Figure 3.7: Partial snapshot of a participant's bookmark showing the distinctive microwave beep pattern in the waveform.

This increased awareness was useful at times for performing household tasks, such as when H2P2 used the system to monitor sound from her washer and dryer and was thus able *"to get my clothes done sooner"* (week 3 survey).

Additionally, H3P1 looked for door open and close sounds in the history view to see when his roommate left, so he could know when to take a shower. Two participants used the display to monitor the well-being of their family members:

> *"[I knew my autistic brother] was pacing around the kitchen [when] I looked at this [display]"* (H3P1)

"It would help [to] recognize if somebody had a fall. If something happened to [my husband who has Parkinson's], I would have no idea. I would just find him on the floor." (H2P2 interview)

Five of six participants also noted how HomeSound provided insight into their *own* behaviors. For example:

> *"I practiced closing and opening the cabinet and monitor the sound waves with the history to improve my ability to be very quiet for [my spouse]. He is a light sleeper…"* (H1P1)
> *"I can turn the kitchen fan off. Earlier my mom/dad used to tell me. [Now,] when they are away, I turn it off on my own."* (H4P1)

However, this increased awareness did not always produce positive reactions or insights. For example,

> *"I was shocked [to learn] how much […] noise I create during meals."* (H2P1, week 1 survey)
> *"I felt embarrassed that this picked up my loud laugh and decided to be careful…"* (H4P1)

Display placement. The physical environment of the home influenced where HomeSound units were placed. In general, participants chose the most active rooms and placed the displays in salient, highly-viewable locations (a shelf or a table). Nevertheless, placement sometimes posed a problem due to visibility or space concerns. For example, to preserve kitchen counter space, H4P1 decided to place the display on top of the fridge but then could not see it. Similarly, participants who opted for a bedroom installation (H1, H3 and H4)

found the screen light disruptive at night, and either covered the displays or put them face down, which decreased their utility. Having the sensing (microphone) and display coupled on a single device also caused issues, some of which reflected the importance of DHH individuals being able to maximize their sightline [58]. For example,

> "*I usually sit over here in my dining room, which [is] a good vantage point for me to see the house. [So,] I placed [the display] here. But then its usefulness was reduced [as] it was far from the kitchen and I wanted the kitchen sounds. So, I moved it to the kitchen. But then, I wasn't able to see it from the dining."* (H1P1)

Decoupling the sensing from the display would also address a desire to have a display where one may not want sensing (*e.g.*, bathroom), as suggested by past work [100] and H3P2:

> *"I wanted a monitor in the bathroom to see what was going on in my home but then I don't want it to [hear] the private bathroom stuff. Can you make the mic and the display separate?"*

Privacy. All occupants accepted the system after learning its privacy-protecting measures but interview responses indicate there could be future privacy concerns. In the initial interview before we explained how HomeSound preserves privacy, three DHH participants had expressed concern about recording conversations, such as: *"Is it recording my voice? Do I have to be concerned about what I am going to say when I am near it"* (H1P1). Surprisingly, no household members beyond the DHH participants expressed privacy concerns to us directly or indirectly. This openness may have been related to its assistive nature. For example, H1P1 said, *"[My hearing spouse] accepted it because it was an assistive technology and he knew this was necessary to help me,"* a sentiment that was also echoed by H4P1:

> *"My mom was concerned when she was cooking, and the system was showing all her cooking activity. But she knew it was important to be there for me to help recognize the sounds."*

However, this notion was not necessarily shared by guests, which included visits from friends and family in two homes:

> *"My friend asked his wife to not hold a conversation near a tablet* […] *Then I explained that this [system] cannot display words and he seemed to be ok with it then. Although I must say he was a little put off initially."* (H3P1)

To mitigate this issue, three participants suggested adding the ability to turn off the recording in a room as needed.

Design suggestions. The most common design suggestions included updates to the visual designs, a notification feature, and automatic sound identification. When asked about the interface design, all participants appreciated the floorplan and history views, finding them easy to learn and use. In contrast, the waveform was seen as too abstract, although it offered an indirect benefit to some participants: H3P1 and H4P1 noted that they had used the waveform to begin recognizing visual patterns of recurring sounds; for example, H4P1 identified the microwave beep.

Four participants also made suggestions for adding features to the history view, such as accessing a stored waveform, daily or weekly summary of information, and the ability to see different timespans. For example, H3P1 wrote about the history view in a week 2 bookmark that, *"I would like to zoom in and see the sound signature [for the microwave beep]."* Finally, H3P1 suggested that different visual designs may be useful for different locations in the home, such as a floorplan view in the office where they sit close to the display, compared to a more generic alert design in the kitchen: *"just some kind of alert that there is a sound."*

Because the visual information was only useful when it was within sight, all participants reported missing useful sounds when they were not close to the display (*e.g.*, door knock, appliance alerts) and requested a notification feature.

> *"I decided if I have to watch the tablet for tea kettle whistle, I may as well watch the tea kettle. Pair with a watch that would vibrate to let you know what is happening."* (H1P1, week 2 text message)
>
> *"If the tablets could flash when some sound occurred, then I could check."* (H2P1)

All participants also expressed the desire for automatic sound classification, so that they did not have to rely so much on context or hearing roommates to guess the source of sounds.

"I would sometimes let the dog out and I always have to make sure to check… Sometimes I would forget that I put him outside and if he wants to come in, he would bark and bark […] It would really help if it tells me the dog was barking." (H2P2)

"I happened to notice [pulses] in the den and bedroom while I was in the dining area. I went to the den to find out what it was, nothing I could ascertain. So, I went to bedroom. Nothing there either. I asked [my hearing spouse] who said it was Siren from outside." (H1P1, week 1 bookmark)

Building on the ability to identify sounds, five participants wanted to know safety-related sounds (*e.g.*, fire alarms) from outside the home.

Study 3 Summary

Participants appreciated HomeSound for its ability to increase self- and home-awareness. They used context such as location and visual cues to identify sounds from basic visualizations, which influenced some daily chores and increased awareness about other occupants' activities. In terms of privacy, house occupants accepted the always-on sound monitoring more than the guests. Finally, the need to constantly monitor the displays, and the lack of automatic sound identification were primary limitations.

3.4 Study 4: Prototype 2 Field Deployment

3.4.1 HomeSound prototype 2

Informed by our experiences with Prototype 1, we extended HomeSound in two ways: first, we added a real-time, deep-learning based sound classification engine to automatically identify and visualize sound events; second we designed and implemented a complementary smartwatch system that provided customizable sound alerts via visual+haptic notifications. We describe both extensions below as well as updates to the IoT display interface.

Sound Classification Engine

To create a robust, real-time sound classification engine, we followed an approach similar to *Ubicoustics* [72], which uses a deep convolutional neural network (CNN) called *VGG16* [41] pre-trained on 8 million YouTube videos [4]. Because VGG16 is developed for video classification, we used transfer learning to adapt

the model for sound classification. For this, similar to [72], we use a large corpus of sound effect libraries—each of which provide a collection of high-quality, pre-labeled sounds. We downloaded 19 common home-related sounds (*e.g.*, dog bark, door knock, speech) from six libraries—BBC [152], Freesound [28], Network Sound [153], UPC [154], TUT [85] and TAU [5]. All sound clips were converted to a single format (44Hz, 16-bit, mono) and silences greater than one second were removed, which resulted in 31.3 hours of recordings. We used the method in Hershey *et al.* [72] to compute input features. Finally, we fine-tuned the model by replacing the last fully connected layer with a fresh layer, retraining on only a subset of sound classes (Table 3.5) to generate per-room classification models.

Table 3.5: List of sounds recognized by our sound classifier.

Kitchen	Bedroom	Living room	Outdoors
Cutlery	Alarm clock	Cat meow	Hammer
Dishwasher	Cough	Dog bark	Drill
Microwave	Snore	Doorbell	Vehicle
Water pour	Door in use	Door in use	
Phone ring	Phone ring	Phone ring	
Speech	Speech	Speech	
Hazard alarm	Hazard alarm	Hazard alarm	
Kettle Whistle		Door knock	

Experimental evaluation. To evaluate our model, we collected our own 'naturalistic' sound dataset. We recorded 16 sound classes from five homes using the same hardware as HomeSound—a Surface Pro 3 with a built-in microphone. For each home, we collected sounds in three rooms (*bedroom*, *kitchen*, *living room*). For each sound class, we recorded five 10-second samples at three distances (*5, 10,* and *15 feet*). We attempted to produce sounds naturally (*e.g.*, using a kettle or running water to wash hands). For certain difficult-to-produce sounds—like a fire alarm—we played snippets of predefined videos on a laptop or phone (33 total videos were used). Because we train per-room classification models, not all sound classes were recorded in each room; Table 3.5 shows the sounds per room (the three outdoor sounds were not recorded for this experimental evaluation). In total, we collected 1,200 recordings (3.3 hrs).

Before testing our model, we also added 20% sound data from other rooms in our test set that our model should ignore (called the "unknown" class). For our evaluation experiment, we classified data collected from each room using the appropriate per-room classification model. Our overall accuracy was 85.9% with small, per-room differences: the average in the living room was 88.5% (*SD*=3.5%) followed by the kitchen (86.4%; *SD*=3.0%) and bedroom (82.5%; *SD*=7.3%). The best performing sounds included cutlery (100%, *SD*=0), door in use (98.7%, *SD*=2.7%), and water pour (96.0%; *SD*=5.3%) and the worst: phone ring (60.0%; *SD*=32.0%), alarm clock (50.7%; *SD*=24.1%), and dishwasher (45.3%; *SD*=7.8%). For poor performing classes, understandable mix-ups occurred: *e.g.,* 24.0% of phone ring sounds were classified as doorbells and 38.4% of alarm clock sounds as a phone ring. Interestingly, accuracy was unaffected by recording distance: at 5ft *avg*=85.5% (*SD*=16.2%), 10ft (85.2%; *SD*=15.9%), and 15ft (87.1%; *SD*=15.1%). We return to classification accuracy and its impact on users in Study 2 Findings and our Discussion.

Implementation. We built the classification engine in Python using *Google TensorFlow* [1], which ran locally on each HomeSound device (to protect occupant privacy): 1 second of microphone data was buffered (44,000 samples) and relevant features extracted and classified. Only the classified sound, classification confidence, loudness, and room location were uploaded to the server (no raw features were transmitted). On the server, all sounds below 50dB or 50% confidence were ignored; the others were broadcast to the web and smartwatch clients.

Smartwatch

To transform HomeSound from a passive awareness system to a proactive one and to eliminate line-of-sight requirements, we designed and implemented a complementary Android-based smartwatch application. The smartwatch displayed a notification along with a vibration alert whenever a classified sound event occurred. The display included sound identify, classification confidence, and room (Figure 3.8). Importantly, each user could customize which sound alerts to receive by clicking on a notification, opening a scroll list, and selecting snooze options (1 min, 5 min, 10 min, 1 hour, 1 day, or forever). In our deployments, we used the *Android Ticwatch E2* watch [155] running *WearOS 2.0*, which communicated with the backend server using WiFi. To enable notifications even when the watch was in a low-power sleep state, we used the firebase messaging service (FCM) for watch-server communication.

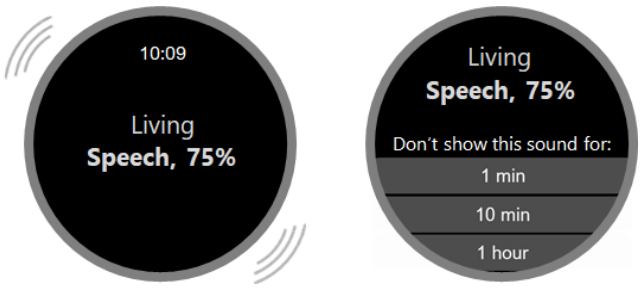


Figure 3.8: HomeSound Prototype 2 smartwatch interface. When a sound occurs, a vibration and a visual notification is received on the watch (left), which when clicked, opens the main app (right) that allows users to snooze the sound.

HomeSound Display Updates

For the HomeSound IoT display, we made three primary changes: first, to incorporate the real-time sound classification engine, we visualized sound identities and their confidence below each circle pulse in the floorplan view and as annotations on the 'sound event' blocks in the history view (Figure 3.9). Second, similar to the smartwatch application, we added a customization menu, which allowed users to select which sounds to show on each display. Finally, we added a pan-and-zoom feature to the history view to increase granularity of sound event visualizations.

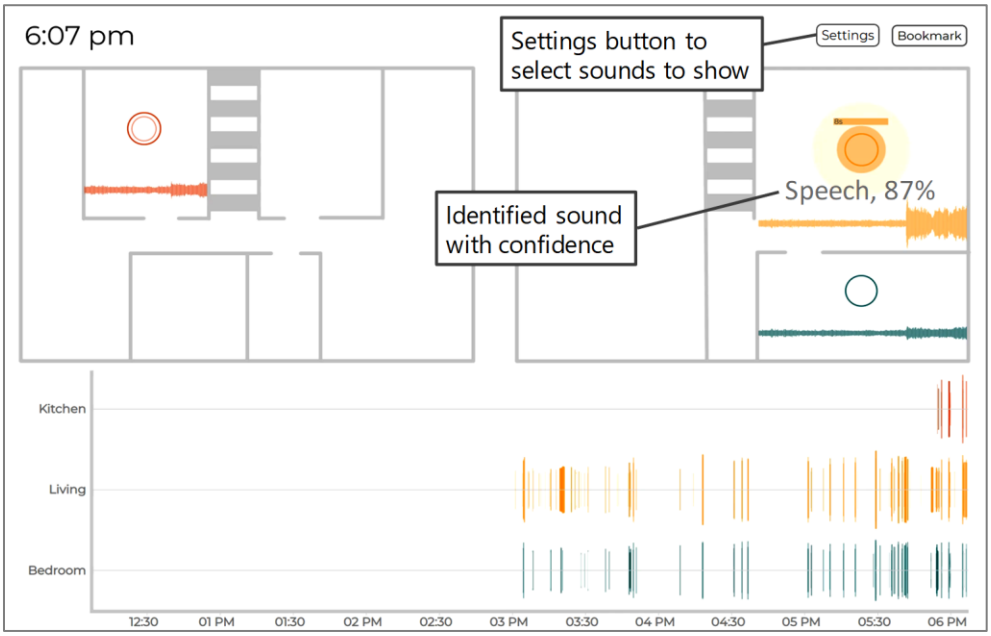


Figure 3.9: HomeSound Prototype 2 interface for displays.

To evaluate Prototype 2, we performed another field deployment in four homes using an adapted Study 3 protocol.

3.4.2 Method

Participants

As before, we recruited DHH participants and other house members through email, social media, and snowball sampling. As an iterative deployment, we did not exclude repeat participants; hence, two of the four homes (H1, H2) were the same as in Study 1 (Table 3.3). Six DHH and two hearing people agreed to participate, whose age averaged 53.1 years old (*SD*=20.9, *range*=21–79). Four DHH participants reported onset of hearing loss as congenital, H5P2 reported at 2 years old and H1P1 reported 3 years old. Two participants used an assistive device: hearing aids.

Procedure

We followed the same process for weekly surveys, and data logging as Study 1 but made slight changes to the initial session and post-trial interview. For the initial session, we made three modifications. First, to generate the room-specific models, participants selected up to eight sounds for each room from the 19-sound list (Table 3.5). Second, to demonstrate the smartwatch app, the researcher produced and snoozed a sound (*e.g.*, speech); participants were also asked to charge the watch at night and wear it throughout the day, including outside the home if desired by connecting to a WiFi source. Third, for the initial interview, new participants responded to the same questions from Study 1 while repeat participants were reminded of their Study 1 responses and asked if anything had changed. Finally, for the post-trial interview, we added two questions on usage and experience with the sound classification and smartwatch app.

Data Analysis

We followed the same Study 1 analysis process with the same two coders. In summary, the final codebook contained 9 level-1 codes, 21 level-2 codes, and 65 level-3 codes. K-alpha was 0.78 (*SD*=0.14; *range*=0.62-0.94) and raw agreement was 91.7% (*SD*=4.3; *range*=85.8-97.2). All disagreements were resolved via consensus.

3.4.3 Findings

We discuss new insights related to Study 1 themes (usage patterns, household tasks, placement, privacy) as well as new emergent themes (cultural differences, playful interactions).

Usage patterns. On average, each home had 3,297.4 sound events/day (*SD*=819.1); 65.9% of them (2174.2, *SD*=525.4) were automatically classified (Table 3.6). Participants filled all surveys, created 41 bookmarks, sent 13 email threads and 32 text messages. In the first week, email and text messages asked about system operation—particularly on how to snooze the smartwatch app or select sounds on the displays, indicating a higher learning curve than Study 1. H6P1 corroborated this: *"at first you have to get used to it. Like a new computer [...] it took time to learn".*

Smartwatch: In general, participants appreciated the watch alerts and wore the watch consistently, except when sleeping, bathing, or going out. Three participants chose to wear the watch outside the home but only wanted to be alerted about urgent home sounds (*e.g.*, a fire alarm) and felt irritated about snoozing other sounds, which indicates a need for location-aware customization. In the home, notifications diminished the need to actively monitor the IoT displays but the 'alert' vibrations could be distracting and overly persistent. Three participants reported using the snooze feature; two others became inured—for example, H2P1 *"would just ignore them when working on my computer"* (interview). One participant removed the watch itself when the noise levels were high:

> *"I had company last Sunday. All of a sudden, it began [vibrating] constantly. I couldn't take away my attention because I didn't want to be rude to my company [and] click [on the app] to snooze different sounds. It was easier to take it off."* (H1P1)

Displays: All displays ran continuously for three weeks, except in H5 where the bedroom display was closed every night for privacy. As smartwatches provided sound alerts, all participants looked at the displays only to recap the events of the day; consequently, for repeat participants, the perceived value of the IoT displays decreased. Only H2P2 and H5P1 described incidents of using the displays immediately after a sound occurred

if one was nearby and in view. The watch also decreased the need to have multiple displays, as corroborated by a repeat participant during the interview:

> *"I check the tablet when I get home from work and see what had happened [...] For the historic information, having one tablet was sufficient. Having multiple tablets was overwhelming"* (H2P2)

Table 3.6: Study 2 Homes, their sizes (in sq. feet), number of displays deployed, most active room, average events logged and identified each day, and total bookmarks by participants

Home	Size	Floors	Displays	Busiest	Events	Identified	Bookmarks
H1	1060	1	4	Den	3647.4	2417.5	12
H2	1740	2	5	Dining	1986.5	932.2	8
H5	1900	2	5	Living	4754.8	3213.6	17
H6	4453	1	4	Kitchen	2801.1	2133.3	4

Household tasks. In some cases, HomeSound helped participants perform household tasks by alerting them to desired sounds in the house (*e.g.*, someone knocking, dog barking, children's shouts). For example,

> *"I was […] working on my laptop, the watch showed my dog was barking [in another room]. I went and corrected my dog right away. This helps me train the dog over time [...] Also, the watch lets me know when the washer is done."* (H2P2, week 2 survey)
>
> *"The first day [when] the contractor would come over for the kitchen remodel. I was sitting close to the door. But the watch vibrated and [displayed] "door knock" and I thought, oh [from] now [on,] I don't have to sit and wait."* (H6P1)

However, system failures resulting from sounds being not supported (*e.g.*, dryer beeps, garbage disposal) or misclassified also negatively affected daily routines. For example, H2P1 said: *"a fan running in the kitchen kept identifying as microwave [....] and I will go and check again and again."* More concerning was when the same sound was only sometimes misclassified. For example, the system correctly identified door knocking initially, so the hearing children of H5P2 assumed that the system would always notify their parents of door knocks, creating issues:

> *"So, somebody was knocking at the door and [my kid] thought that [...] my watch will tell me and didn't bother to come up to me [...] And the guy was knocking, knocking, and finally [my kid] comes up to me and [signs], why are you not opening the door!?"*

Interestingly, four participants found creative ways to compensate for some misclassifications using the help of house members, animals, or context:

> *"Someone knocked at the door [….] and [HomeSound] was not recognizing any of it. But [my dog] barked, and the watch alerted me to "dog bark", so I went and looked."* (H2P2)
>
> *"I know microwave cannot run in my bedroom, so I ignored it."* (H6P1, to a follow-up question on her bookmark)
>
> *"It said hazard alarm, but no lights were flashing. [My hearing spouse] confirmed it was from outside."* (H2P1)

Playful interactions. Beyond household tasks, all occupants reported instances of deliberately initiating actions to record and explore their sound "footprint". For example, H1P1 would sometimes *"clap, hoot, or talk at the system for seeing them later in time."* Though this behavior mostly occurred in the first week, it resurfaced when guests arrived (in H2 and H6). In addition, H5P1 reported that his two children liked seeing the annotations in history view, and they would *"scream at [the displays] to see bubbles going up, down [pulsating in floorplan]"* or would "*play a variety of sounds*" to see how well the system would perform.

System improvements. Participants offered concrete ways to improve the smartwatch, display, and sound classification.

Smartwatch: To reduce the constant smartwatch vibration, five participants suggested alerting about a sound only once:

> *"If somebody is talking, it should tell me once, and that's it. This alerts me every second, and I [snooze for some time] and it comes back again. I can't [snooze] indefinitely, because what if somebody is actually calling me."* (H1P1)

He added: *"also, if it showed speech, I would want to know [who] is talking."* This need to provide more details on the speech sound (*e.g.*, speaker id, tone) was highlighted by three other participants as well.

Display: Because of the smartwatch, participants used the displays mainly to view past sounds and suggested enlarging the history view (N=5); four wanted to remove the floorplan and add the floorplan's characteristics (loudness, duration, pitch, color) to the smartwatch app (N=5) or the history view (N=2). For example, while sketching a new design during the interview, H1P1 said:

> *"[No] need to have the layout (floorplan) […] Just show history [...] Make it larger to fit the entire screen [sketching]. Right now, the color codes are used to identify a room. But perhaps color codes could be used to [distinguish] sounds in the history".*

Classification: To compensate for sound classification issues, participants suggested three technical improvements. First, three participants wanted more precise localization in areas with many sound producing appliances (*e.g.*, kitchen) so they can identify sounds from their location. Second, three participants suggested dynamically adjusting the microphone sensitivity to increase feedback utility:

> *"Before remodeling the kitchen (in second week of deployment) we had a porcelain sink. We have stainless steel sink now. The water is quieter [in porcelain sink] but when it hits the stainless-steel sink [the sound] is amplified. So, too loud now and [...] I had to move the [display] a little farther."* (H1P1)

Finally, participants suggested fine tuning the system (N=3) to the sounds of their home. For example, H6P1 said:

"better to record some sounds myself […] I would prefer knowing [my spouse's] speech rather than knowing everybody else's"

Placement. The addition of the smartwatch and sound classification feature changed how occupants positioned the IoT displays in their home. While there was a decreased need for line-of-sight, participants felt that HomeSound needed to be close enough to sound activity to accurately detect and classify sound events. Consequently, three homes moved the displays closer or farther from sound sources, during which issues with space emerged:

"But then my kitchen is small… there's [only] so far I could move it. So, I placed it [a little outside the kitchen]" (H1P1)

And H5P2 said in the interview:

"There was nothing [no shelf space] closer to the front door, so I had to bring a table but then the door wouldn't [fully] open"

Another theme related to how sounds propagate through a home, which could be confusing or raise privacy concerns:

"I saw hammering in multiple rooms which surprised me […] But then it occurred to me that these three rooms are all closer to the street, and so that must be the loud construction noise [from outside]." (H6P1, week 1 bookmark)

"Even after turning [the bedroom tablet] off [at night], we were concerned whether the tablets in other rooms [close to our room] would pick up the signal and the kids can see from downstairs [living room tablet]. The children might think it's weird to be having many sounds at this time of the night…" (H5P2)

Cultural differences. Though all participants reported that HomeSound helped them in some way, subtle differences emerged between Deaf (H2, H5) and hard of hearing (H1, H6) households, potentially related to cultural context (*e.g.*, Deaf people tend to rely less on sounds than hard of hearing people [13]). Indeed, the three Deaf participants expressed that, apart from important cues (*e.g.*, doorbell, alarm), other information was "nice to know" rather than "need to have":

"I've been Deaf since I was an infant, so I am used to life without sound. It is not really a big concern unless it is an emergency… I think the system could be better for people who became deaf later in life or hard of hearing but not necessarily for someone like myself." (H2P1).

However, Deaf individuals may find the sound information more valuable with longer-term exposure. For example:

"It was surprising that I'm making many noises such as closing the door, cabinets, talking to my dog, putting food and flatware on the counter (cutlery noises). […] I felt awkward subjecting my (hearing) kids to all this noise and tried to change." (H5P2)

Privacy. Surprisingly, though with Prototype 2, participants received constant notifications about sound events (with more information), privacy concerns did not change from Study 1. Participants and their household members reported no issues except in the case of guests (H2, H6) who were more curious than suspicious, and the issue of display placement with children (H5) (both are detailed above).

Study 4 Summary

The smartwatch and sound classifications diminished the importance on the IoT displays while increasing general sound awareness. Key concerns included system failures, unpredictable sound classification errors, and overly persistent watch vibrations.

3.5 Discussion

Our studies confirm past work with DHH participants on sounds of interest in the home [2,14], as well as identify new preferences for home-based sound awareness (*e.g.*, integrate with existing infrastructure), and design considerations that manifest in unique ways in the home setting (*e.g.*, privacy, the utility of displaying sounds in a home layout).

We also presented the design and field evaluation of the first IoT-based sound awareness system for DHH users. Some of our results (*e.g.*, feelings about privacy, information overload) contextualize the prior findings from lab-based evaluations [13,53,80] but with higher ecological validity. We also report on new findings that are only possible via a real-world field deployment, such as usage patterns over time, reactions to varying

locations of displays at home, and influence on social dynamics. Below, we discuss further implications of our findings and opportunities for future work.

3.5.1 Support for Domestic Processes

Work in domestic computing supports that technologies should interweave into the living processes of the household [6,22]. Similarly, sound awareness technologies need to be contextually aware of home activities, and selectively display sounds. Several contextual parameters arose in our studies including the time of the day (*e.g.*, night *vs.* day), user's current activity (*e.g.*, working *vs.* idle), amount of activity in the home (*e.g.*, a large guest party *vs.* quiet), and the user's location (*e.g.*, kitchen *vs.* bedroom). Home-based sound awareness systems should be designed with these factors in mind, and further research with functional systems will help to understand how best to support these different contexts. Future technologies should also integrate with existing home adaptations of DHH users (*e.g.* vibratory wake-up alarms, R3) and other home systems (*e.g.*, security camera, R9).

3.5.2 Self-Awareness

An unintended effect of the system was increased self-awareness, leading to adaptations in behavior. Thus, future work should explore how the system can better support these feedback-based adaptations, for example, by showing volume graphs and notifications of personal sounds above a certain volume. Past work in tactile-based sound awareness examined wrist-worn devices for regulating personal voice levels [123], but this work could be extended to visual displays and smartwatches in the home context. Importantly, there can be negative implications to any system that explicitly or even inadvertently encourages users to change their behavior. For example, while most feedback was received positively, the visualizations of loud noises created by participants were sometimes associated with feelings of embarrassment.

3.5.3 Privacy

The home is a complex and evolving space shared with other family members and guests, hence concerns of privacy arise [5]. Introduction of a sound awareness system into a space inhabited by deaf or hard of hearing occupants may change the notion of privacy—visual privacy means occluding line of sight, whereas sound can be sensed remotely. Investigating the implications of this difference will be important, particularly with

consideration of cultural context. Indeed, our findings suggest that deaf/Deaf participants may have felt differently about intimate sounds than hard of hearing participants.

Attention should also be given to the non-DHH members of the household. In our studies, these hearing members seemed to accept the system because it was perceived as 'assistive' [106], through further exploration is needed.

Another dimension of privacy is about unwanted access to historical sound information. Placing the display out of view of the "public" areas of the house may mitigate this concern, but would also reduce the ease of accessing the information. In general, future work should consider: who should be able to view the sound information and what inferences can be drawn about human activity from the visual representations?

3.5.4 Handling Uncertainty

Home technologies need to gain the user's trust [6]. Similar to Mielke *et al.* [17], we found that false negatives (i.e., not showing an occurred sound) were deemed more tolerable than false positives (i.e., showing a sound that did not occur), and that misattribution (i.e., showing a wrong sound) would result in losing trust in the system. Our evaluation also suggests a more concerning case: when these errors are unpredictable—which can cause users' expectations and system behavior to be mismatched.

The system misclassification issues suggest a need to further improve system accuracy or, at the very least, mitigate the potential downsides of inaccurate or unpredictable behavior. One possibility—as we explore in Chapter 5—is to employ a customization approach that allows participants to train the system for their home, though this training may be tedious and difficult if the sound itself is inaccessible to the user. While we conveyed classification confidence to users, there may be opportunities to adapt how sound information is displayed based on confidence, such as displaying more ambiguous information (*e.g.*, a sound occurred), a possible list of sounds, or a general indication of the sound (*e.g.*, "an alarm-like" sound) when confidence is low (as opposed to simply choosing not to show low confidence sounds at all, as in our design).

3.5.5 Form Factor and UI

In our formative studies (Study 1 and 2), participants preferred the floorplan prototype the most because, as reflected in a past work [14], it visualizes sounds in a spatial layout. Our new list prototype, which displays

a temporal history, was also well received. In contrast to the prior work [80], the waveform prototype was not preferred as a standalone design because of low information gain, but participants found value in using it in combination with the list and spatial layout prototype. Thus, in our field evaluations, we investigated a design that combines the three prototypes together, which our participants appreciated in general, but the specific reactions differed between the two studies (Study 3 and 4). While the floorplan was appreciated in Study 3 as an indicator of current sounds and their location, the addition of the smartwatch alerts reduced this utility. Thus, participants suggested removing the floorplan and supplementing the sound information (*e.g.*, loudness, duration) on the smartwatch. Future work should further investigate how to balance this information between the small smartwatch face and tablets.

Finally, though we did not conduct an evaluation, half of participants in Study 1 found promise in ambient displays (*i.e.*, illuminating sound sources within a room), which should be further explored.

3.5.6 A Note on Participant Diversity

Our participants identified as Deaf, deaf or hard of hearing. Deaf (capital 'D') refers to people who belong to a Deaf culture with distinct norms and practices, whereas deaf (small 'd') and hard of hearing indicate people with hearing loss who may or may not identify with Deaf culture [4]. Further, the terms 'deaf' and 'hard of hearing' can refer to audiological differences in hearing level. Despite these differences, disability occurs on a spectrum and these groups have synergetic access needs and preferences (as echoed by the majority of our findings as well as past work, *e.g.*, [2]). Recruiting a wide range of participants at this initial stage of research allowed us to explore solutions that would work for a diversity of users. Future work should also focus on the needs of specific subgroups within the DHH population.

3.5.7 Limitations

First, our field study findings are based on two three-week deployments but with only four homes. Future work should include a larger and more varied set of households, which could yield additional insights, especially related to social dynamics. Second, though our controlled evaluation showed acceptable overall accuracy, we do not have quantitative data on how our sound classification system worked in practice; instead, we only have participants self-reported perceptions of the system. Finally, while our designs were informed by prior work [53], other visual and haptic representations are possible and should be explored.

3.6 Chapter Summary

In this work, we explored the DHH people's preferences for future in-home sound awareness technology and solicited their reactions to an in-home sound awareness system. Our findings demonstrate value, especially with regards to feelings of increased awareness amongst our DHH participants but also uncover important issues related to privacy, contextual awareness, social dynamics, and handling AI uncertainty. Our work has implications for future 'smarthome' displays such as the Google Home or Amazon Echo Show.

Chapter 4: Investigating Sound Awareness on Portable Devices

While the previous chapter investigated sound awareness in the home, our findings reveal that DHH people also wanted sound awareness outside the homes as well. Consequently, we turned to examining portable solutions for providing sound feedback, and chose smartwatch, which according to our recent survey with 201 DHH participants [25] is the most preferred device for non-speech sound awareness. Indeed, smartwatch has integrated support for both visual and vibrational modalities and can offer glanceable, always available, and private sound feedback in multiple contexts.

Most prior work in wearable sound awareness, however, has focused on smartphones [13,89,119], head-mounted displays [35,43,49], and custom wearable devices [61,102] that provide limited information (*e.g.*, loudness) through a single modality (*e.g.*, vision). A few Wizard-of-Oz studies have explored using visual and vibrational feedback on smartwatches for sound awareness [33,88,89]; however, the evaluations of the prototypes were preliminary. One exception includes Goodman *et al.* [33], who conducted a Wizard-of-Oz evaluation of smartwatch-based designs, gathering user reactions in different audio contexts (a student lounge, a bus stop, and a cafe). However, this work was intentionally formative with no functioning implementations.

Furthermore, recent deep-learning research has investigated multi-class sound classification models, including for DHH users [54,119]. For example, Jain *et al.* [54] used deep convolutional neural networks to classify sounds in the homes of DHH users, achieving an overall accuracy of 85.9%. While accurate, these cloud or laptop-based models utilize a high memory and processing power and are unsuitable for low-resource portable devices.

Building on the above research, in this chapter, we present two smartwatch-based studies. First, we quantitatively examine four state-of-the-art low-resource deep learning models for sound classifications: *MobileNet* [44], *Inception* [125], *ResNet-lite* [126], and a quantized version of *HomeSound* [54], 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 tradeoffs in computational and network requirements, power efficiency, data privacy, and latency. While direct comparison to prior work is challenging, our experiments show that the best classification model (VGG-lite) performed similarly to the state of the art for non-portable devices while requiring substantially less memory (~1/3rd). We also observed a strict accuracy-latency trade-off: the most accurate model was also the slowest (*avg. accuracy*=81.2%, *SD*=5.8%; *avg. latency*=3397ms, *SD*=42ms). Finally, we found that the two phone-based architectures (*watch+phone* and *watch+phone+cloud*) outperformed the watch-centric designs (*watch-only*, *watch+cloud*) in terms of CPU, memory, battery usage, and end-to-end latency.

To complement these quantitative experiments, we built and conducted a qualitative lab evaluation of a smartwatch-based sound awareness app, called *SoundWatch* (Figure 4.1), 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 semi-structured interview about their experiences, their views regarding accuracy-latency tradeoffs and privacy, and ideas and concerns for future wearable sound awareness technology. We found that all participants generally appreciated SoundWatch across all three contexts, reaffirming past sound awareness work [25,33]. However, misclassifications were concerning, especially outdoors due to background noise. For accuracy-latency tradeoffs, participants wanted minimum delay for urgent sounds (*e.g.*, car honk, fire alarms)—to take any required action—but maximum accuracy for non-urgent sounds (*e.g.*, speech, background noise) to not be unnecessarily disturbed. Finally, participants selected *watch+phone* as the most preferred architecture because of privacy concerns with the cloud, versatility (no Internet connection required), and speed (*watch+phone* classified faster than *watch* only).

Figure 4.1: *SoundWatch* uses a deep-CNN based sound classifier to classify and provide feedback about environmental sounds on a smartwatch in *real-time*. Images show different use cases of the app and one of the four architectures we built (*watch+phone*).

4

4.1 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 from previous work [33,54], SoundWatch informs users about three key sound properties: *sound identity*, *loudness*, and *time of occurrence* through customizable sound alerts using visual and vibrational feedback (Figures 1 and 3). We use a deep learning-based sound classification engine (running on either the watch or on the paired phone or cloud) to continually sense and process sound events in real-time. Below, we describe our sound classification engine, our privacy-preserving sound sensing pipeline, system architectures, and implementation. The SoundWatch system code is open sourced on GitHub: https://github.com/makeabilitylab/SoundWatch and our app is available on Google Play: https://play.google.com/store/apps/details?id=com.wearable.sound (see a demo video).

4.1.1 Sound Classification Engine

To create a robust, real-time sound classification engine, we followed an approach similar to *HomeSound* [54], which uses transfer learning to adapt a deep CNN-based image classification model (VGG) for sound classification. We downloaded four recently released (in Jan 2020 [156]) TensorFlow-based image-classification models for small devices: *MobileNet*, 3.4MB [44], *Inception*, 41MB [125], *ResNet-lite*, 178.3MB [126], and a quantized version of the model used in HomeSound [54], which we call *VGG-lite*,

281.8MB. Since the size of four models differ considerably, we hypothesized that they would offer different tradeoffs in terms of accuracy and latency.

To perform transfer learning, similar to Jain *et al.* [54], we used a large corpus of sound effect libraries—each of which provide a collection of high-quality, pre-labeled sounds. We downloaded 20 common sounds preferred by DHH people (*e.g.*, dog bark, door knock, speech) [13,53] from six libraries—BBC [152], Freesound [28], Network Sound [153], UPC [154], TUT [85] and TAU [5]. All sound clips were converted to a single format (16KHz, 16-bit, mono) and silences greater than one second were removed, which resulted in 35.6 hours of recordings. We then divided the sounds into three categories based on prior work [13,80]: high priority (containing the 3 most desired sounds by DHH people), medium-priority sounds (10 sounds), and all sounds (20 sounds) (see Table 1). We used the method in Hershey *et al.* [41] to compute the *log mel-spectrogram* features in each category, which were then fed to the four models, generating three models of each architecture (12 in total).

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

4.1.2 Sound Sensing Pipeline

For always-listening apps, privacy is a key concern. While 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 non-reconstructable mel-spectrogram features. While the uploaded features can be used to identify the kind of activity a user is engaged in (*e.g.*, speaking, cooking), conversational information is not retrievable. For signal processing, we take a sliding window approach: the watch samples the microphone at 16KHz and segments data into 1-second 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 to the user, the others are ignored.

4.1.3 System Architectures

We implemented four device architectures for SoundWatch: *watch-only*, *watch+phone*, *watch+cloud*, and *watch+phone+cloud* (Figure 4.2). 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 classification. For only the cloud-based architectures, to protect user privacy, non-reconstructable sound features are computed before being sent to the cloud—that is, on the watch (*watch+cloud*) or on the phone (*watch+phone+cloud*). We use Bluetooth Low Energy (BLE) for watch-phone communication and WiFi or a cellular network for watch-cloud or phone-cloud communication.

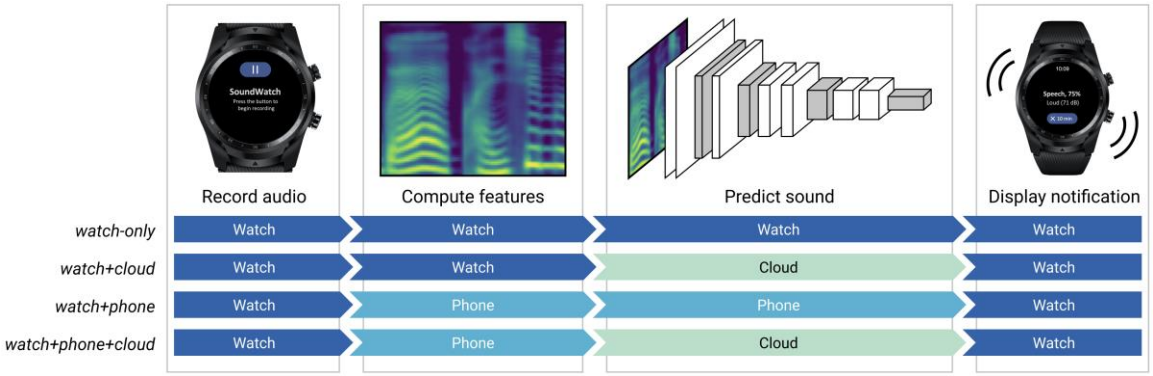


Figure 4.2: A diagram of the four SoundWatch architectures and a breakdown of their sensing pipelines. Block widths are for illustration only and are not indicative of actual computation time.

4.1.4 User Interface

To increase 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 (Figure 4.3). 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 hour, 1 day, or forever). Additionally, the user can select which sounds to receive alerts for by using the paired phone app, which displays a customization menu (Figure 4.3d). While future versions should run as an always-available service in Android, currently the app must be explicitly opened on the watch to run (Figure 4.3a). Once the app is opened, it continuously runs in the background.

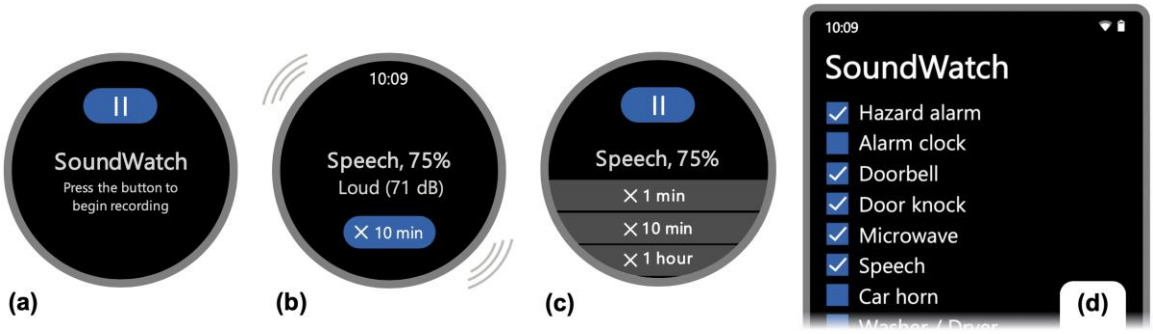


Figure 4.3: The SoundWatch user interface showing the opening screen with a button to begin recording the audio for classification (a), the notification screen with a "10-min" mute button (b), and the main app screen with more mute options (c). (d) shows a partial view of the paired phone app to customize the list of enabled sounds.

4.2 System Evaluation

To assess the performance of our SoundWatch system, we perform two sets of evaluations: (1) a comparison of the four state-of-the-art sound classification models for embedded 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 *Android Ticwatch Pro* watch (4×1.2GHz, 1GB RAM) [157] and the *Honor 7x* Android phone (8×2GHz, 3GB RAM) [158]. For emulating the cloud, we used an Intel i7 desktop running Windows 10.

4.2.1 Model Comparison

To determine how different models perform on the watch and phone, we trained and evaluated the classification accuracy and speed of our four model architectures. To compare with prior approaches in sound classification, we also evaluated the full-VGG model (281.8MB) on a non-portable device. Below we detail the experiments and results.

Accuracy

To calculate the "*in-the-wild"* inference accuracy [159] of the models, we collected our own 'naturalistic' sound dataset similar to *HomeSound* [54]. We recorded 20 sound classes from nine locations (three homes, three offices, three outdoors) using the same hardware as SoundWatch: the 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). We attempted to produce sounds naturally (*e.g.*, using a microwave or opening the door). For certain difficult-to-produce sounds—like a fire alarm—we played snippets of predefined videos on a laptop or phone with external speakers (54 total videos were used). In total, we collected 540 recordings (~1.5 hours).

Before testing our model, we divided our recordings into the three categories (all sounds, high priority, medium priority) similar to our training set (Table 1). For the medium and high priority testsets, 20% of the sound data that we added was from excluded categories that our models should ignore (called the "unknown" class). For example, 20% of the high priority testset included recordings from outside of the three high priority sound classes (fire/smoke alarm, alarm clock, door knock).

For this experiment, we classified data in each category using the models. The results are shown in Figure 4.4. 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 *post hoc* 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 sounds decreased from all (20 sounds) to medium (10 sounds) and high priority (3 sounds). For example, if we only classify the three highest-priority sounds, our average accuracies increase from 81.2% (*SD*=5.8%) to 97.6% (*SD*=1.7%) for VGG-lite and from 65.1% (*SD*=10.7%) to 78.1% (*SD*=11.9%) for ResNet-lite. Finally, in analyzing performance as a function of location context, home and office outperformed outdoors for all models. With VGG-lite, for example, average accuracies were 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 sound recordings may have incurred interference due to the background noise.

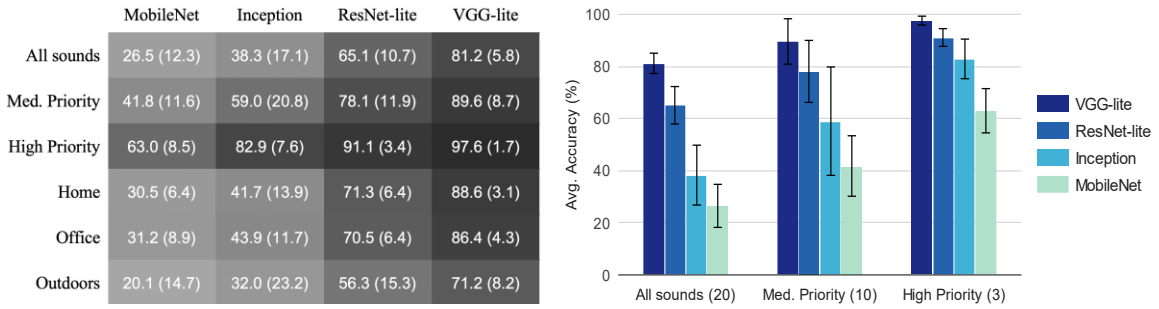


Figure 4.4: Average accuracy (and SD) of the four models for three sound categories and three contexts. Error bars in the graph show 95% confidence intervals.

To further assess model performance, we computed a confusion matrix for medium-priority sounds, which helps highlight inter-class errors (Figure 4.5). While per-class accuracies varied across models *microwave*, *door knock*, and *washer/dryer* were consistently the best performing classes with VGG-lite achieving average accuracies 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 poorer performing classes, understandable mix-ups occurred—for example, alarm clocks and phones rings, which are similar sounding, were commonly confused.

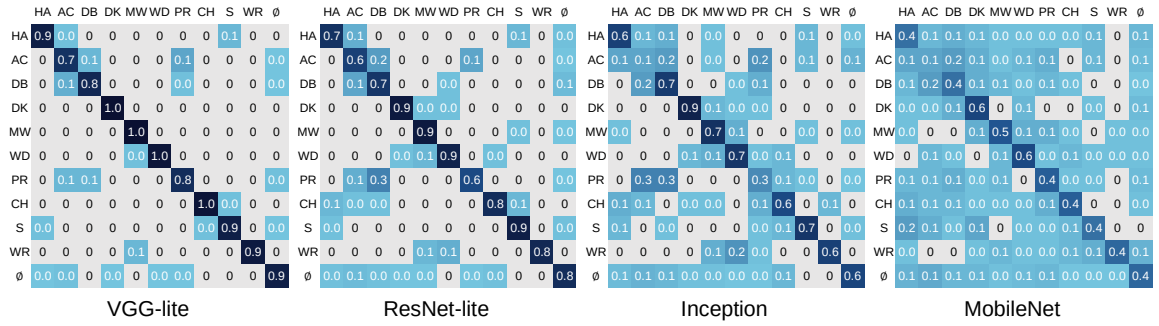


Figure 4.5: Confusion matrices for the four models when classifying 10 sounds in the medium-priority category. Darker blue indicates higher accuracy. *HA*=Hazard Alarm, *AC*=Alarm Clock, *DB*=Doorbell, *DK*=Door Knock, *MW*=Microwave, *WD*=Washer/Dryer, *PR*=Phone Ringing, *CH*=Car Horn, S=*Siren*, *WR*=Water Running, Ø=Unknown.

Latency

In addition to accuracy, the *speed* with which a model performs classifications is crucial to achieving a real-time sound identification system. To evaluate model latency, we measured the time required to classify sounds from the input features on both the watch and the phone. We wrote a script to loop through the sound

recordings in our dataset for three hours (1080 sounds) and measured the time taken for each classification. Understandably, the latency increased with the model size: the smallest model, MobileNet, performed the fastest on both devices (*avg. latency* on watch: 256ms, *SD*=17ms; phone: 52ms, *SD*=8ms), followed by Inception (*avg. latency* on watch: 466ms, *SD*=15ms; phone: 94ms, *SD*=4ms), and ResNet-lite (*avg. latency* on watch: 1615ms, *SD*=30ms; phone: 292ms, *SD*=13ms). VGG-lite, the largest model, was the slowest (*avg. latency* on watch: 3397ms, *SD*=42ms; phone: 610ms, *SD*=15ms).

In summary, for phone and watch models, we observed a strict accuracy-latency tradeoff—for example, the most accurate model VGG-lite (*avg. accuracy*=81.2%, *SD*=5.8%) was the slowest (avg. latency on watch: 3397ms, *SD*=42ms). 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: 1615ms, *SD*=30ms).

Cloud model (VGG-16)

To attempt comparison with past work, we also evaluated the performance of the full VGG model [72] on the cloud. On average, the 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%). This result is promising because our VGG-lite model is more than three times smaller than VGG (281.8MB *vs.* 845.5MB). However, the full model on the cloud performed much faster (*avg. latency*=80ms, *SD*=5ms) than our models on phone or watch.

4.2.2 Architecture Evaluation

Besides model evaluation, we also compared the performance of four different architecture designs for the SoundWatch system: *watch-only*, *watch+phone*, *watch+cloud*, and *watch+phone+cloud*. These architectures differ in terms of where classification computations occur, battery usage, classification speed, network requirements, and privacy—which impacts both technical performance and usability.

For our architecture evaluation, we used the most accurate model on the watch and phone: VGG-lite; the cloud used the full VGG model. Informed by prior work [39,62,82], we measured CPU, memory, and network usage, end-to-end latency, and battery consumption. For the test, we used a script running on a laptop that looped through the sound recordings for three hours 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 (*e.g.*, see [82]).

To determine CPU, memory, and network usage, we used *Android Profiler* [160], a commonly used tool in the literature [45]. For the battery, we used *Battery Historian* [161]. Finally, to determine end-to-end latency, we measured the elapsed time (in milliseconds) between the start of the sound recording window to when the notification is shown. Below, we detail the results.

CPU Utilization

Minimizing CPU utilization is crucial to maximizing the smartwatch's battery performance and lowering the impact on other running apps. Our results for CPU usage on the watch and phone are shown in Figure 4.6a. As expected, the watch's CPU utilization was lowest when classifications were performed on the phone (*watch+phone*; *avg.*=22.3%, *SD*=11.5%, *max*=42.3%) or in the cloud (*watch+phone+cloud*; *avg.*=23.0%, *SD*=10.8%, *max*=39.8%). Here, the watch is used only for *recording* sounds, *transmitting* data via Bluetooth, and *displaying* sound feedback. For *watch+cloud*, the watch is computing the sound features and communicating directly with the cloud via WiFi for classification, which resulted in significantly higher CPU utilization (*avg.*=51.1%, *SD*=14.9%, *max*=76.1%). Finally, if the entire classification model runs on the watch directly, the CPU utilization is nearly maxed out (*avg.*=99.0%, *SD*=2.1%, *max*=100%) and is thus not practical for real-world use.

Memory usage

A smartwatch app must also be memory efficient. We found that the memory usage was primarily dependent 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 (Figure 4.6b). This indicates that running a large model like VGG-lite on the watch could exceed the memory capacity of some modern watches (*e.g.*, [155]). The other app processes (*e.g.*, UI, computing features, network) required less than 50MB of memory.

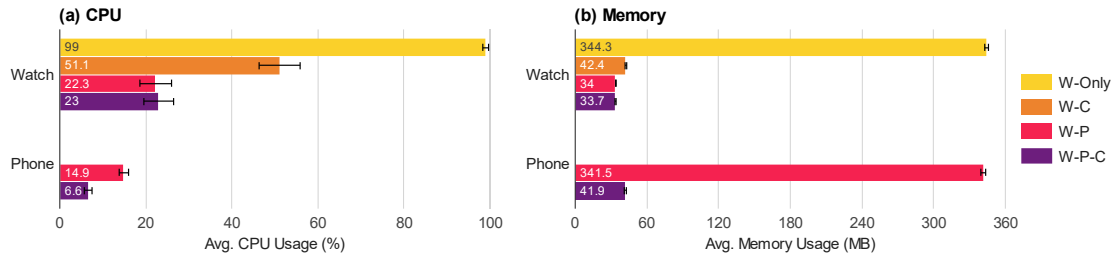


Figure 4.6: Average CPU (a) and memory (b) usage of the four architectures (using the VGG-lite model). Error bars show 95% confidence intervals.

Network usage

Having a low network requirement increases the portability of an app, especially for low-signal areas. Additionally, some users may feel uncomfortable with their data being uploaded to the cloud, even with privacy and security measures in place. For our cloud-based architectures, we found minimal network consumption: for *watch+cloud*, the average 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); both are negligible compared to the network bandwidth of modern IoT devices. The non-cloud architectures used no network bandwidth as they perform all classifications locally on the device(s): either the *watch* or *watch+phone*.

Battery consumption

A fully mobile app needs to be energy efficient. We measured the battery drain from full charge until 30% (Figure 4.7). First considering the watch-based architectures, the *watch-only* architecture used a large amount of battery: 30% at 3.3 hours, a 6.3x increase over the baseline (without our app). Within the remaining three architectures, both *watch+phone* (30% at 15.2 hours, 1.4x over baseline) and *watch+phone+cloud* (30% at 16.1 hours, 1.3x over baseline) were more efficient than *watch+cloud* (30% at 12.5 hours, 1.7x over baseline), because the latter used WiFi which is less energy efficient than BLE [115].

Similar trends were observed on the phone; however, running the model on the phone (*watch+phone*) was still tolerable (1.3x over baseline) as compared to the watch (6.3x over baseline). In summary, we expect that the watch-only architecture would be impractical for daily use, while the other architectures are usable.

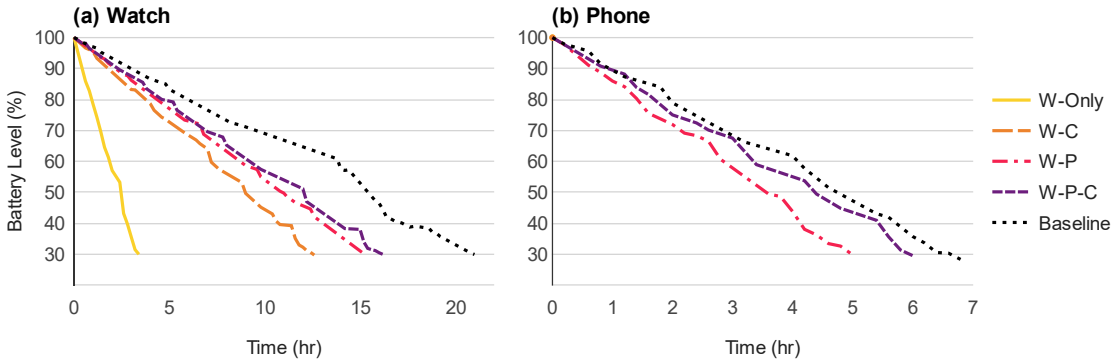


Figure 4.7: Battery level over time for watch (a) and phone (b) for the four architectures. *W-only*=watch only, *W-C*=watch+cloud, *W-P*=watch+phone, *W-P-C*=watch+phone+cloud. Baseline represents the case without the SoundWatch app running

End-to-end latency

Finally, a real-time sound awareness feedback system needs to be performant. Figure 4.8 shows a computational breakdown of end-to-end latency, that is, the total time spent in obtaining a notification for a produced sound. On average, *watch+phone+cloud* performed the fastest (*avg. latency*=1.8s, *SD*=0.2s). This was followed by *watch+phone* (*avg.*=2.2s, *SD*=0.1s), which needed more time for running the model on the phone (vs. cloud), and *watch+cloud* (*avg.*=2.4s, *SD*=0.0s) which required more time to compute features on the watch (vs. phone in *watch+phone+cloud*). As expected, *watch-only* was significantly slower (*avg.*=5.9s, *SD*=0.1s) and is thus, currently unusable (though future smartwatch generations will be more capable). In summary, except for watch-only, all architectures had a latency of ~2s; we evaluate whether this is acceptable in the user study.

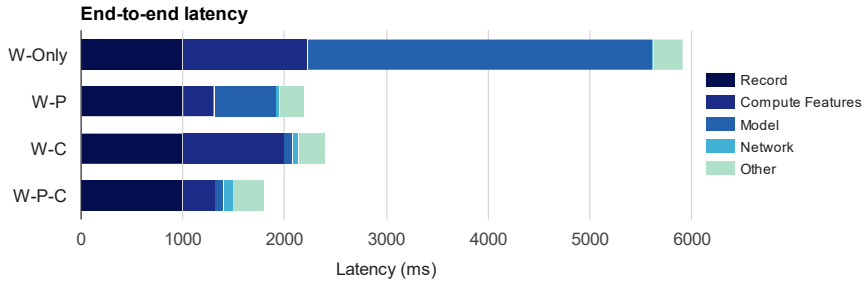


Figure 4.8: Breakdown of end-to-end latency for the four architectures.

Summary

Overall, we found that *watch+phone* and *watch+phone+cloud* outperformed the *watch+cloud* architecture for all system parameters. Additionally, the *watch-only* architecture was impractical for real-life use due to high CPU, memory, and battery usage, as well as a large end-to-end latency. Within the phone-based architectures, the *watch+phone+cloud* performed better than the *watch+phone*.

4.3 User Study

To gather qualitative feedback on our system results and general reactions to smartwatch-based sound awareness in multiple contexts, we performed a lab and campus walkthrough evaluation of our SoundWatch app with eight DHH participants. SoundWatch is designed to support all four device architectures and can be switched between them; however, based on our system experiments above, we used the best performing architecture (*watch+phone*) and model (VGG-lite) for the user study.

4.3.1 Participants

We recruited eight DHH participants (3 women, 3 men, 2 non-binary) using email, social media and snowball sampling (Table 2). 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 reported using hearing devices: three participants 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 reported fluency with reading English (5/5 on rating scale, 5 is best). Participants received $40 as compensation.

Table 2: Demographics of the DHH participants.

ID	Age	Gender	Identity	Hearing loss	Onset age	Hearing device
P1	31	Male	hard of hearing	Moderate	Birth	Hearing aids
P2	26	Female	deaf	Profound	1 year	Cochlear implants
P3	20	Non-binary	deaf	Profound	Birth	Cochlear implants
P4	20	Female	hard of hearing	Severe	Birth	Both
P5	57	Male	deaf	Severe	Birth	Both

P6	23	Female	Deaf	Profound	Birth	Both
P7	38	Non-binary	Deaf	Severe	Birth	None
P8	63	Male	Deaf	Profound	Birth	Cochlear implants

4.3.2 Procedure

The in-person procedure took place on a university campus and lasted up to 90 minutes. Sessions were led by the first 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. Instructions and interview questions were presented visually on an iPad (see supplementary materials), while responses and follow-up discussion were spoken or translated to/from ASL. The session began with a demographic and background questionnaire, followed by a three-part protocol, the first and third of which took place in a quiet conference room:

Part 1: Introduction of SoundWatch prototype (5-10 mins)

In the first phase, we asked about general thoughts on using smartwatches for sound awareness. Participants were then asked to wear the watch running SoundWatch. To demonstrate the app, a researcher made three example sounds (speech, door knock, and phone ring) while explaining the watch and the phone UI. Participants were also encouraged to make their own sounds such as by speaking or knocking to examine SoundWatch's behavior.

Part 2: Campus walk (20-25 mins)

For Part 2, the researcher and the participant (with the watch and phone) visited three locations on campus in a randomized order: (1) a home-like location (lounge of our building), (2) an office-like location (a grad student office), and (3) an outdoor location (a bus stop). These locations enabled the participants to experience SoundWatch in different contexts and soundscapes. In each location, participants used the watch naturally for about five minutes (*e.g.*, by sitting on a chair in an office, or walking and conversing outdoors). In locations with insufficient sound activity (*e.g.*, if the lounge was empty on weekends), the researcher produced some sounds (*e.g.*, by running the microwave, or washing hands). Participants were also encouraged to use the sound customization options (mute on watch and checklist on phone) if they desired.

Before exiting each location, participants filled a short feedback form to rate their experience on a 5-point scale and document any open-ended comments.

Part 3: Post-trial interview (45-50 mins)

After completing the three locations, participants returned to the lab for Part 3. Here, we conducted a semi-structured interview inquiring about the participant's overall experience and perceptions of SoundWatch across the different locations, reactions to the UI, privacy concerns, and future design ideas. We then transitioned to asking about specific technical considerations, including accuracy-latency tradeoffs and the four possible SoundWatch architectures. For accuracy-latency, we explained the concept and then asked about their expectations for minimum accuracy, maximum delay, and whether their perspectives changed based on sound type (*e.g.*, urgent *vs.* non-urgent sounds) or context (*e.g.*, home, office). To help discuss the four SoundWatch architectures—and to more easily allow our participants to understand and track differences—we prepared a chart (see supplementary materials) enumerating key characteristics such as battery or network usage and 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 to specify reasons for their rating.

4.3.3 Data Analysis

The interview transcripts and the in-situ form responses were analyzed using an iterative coding approach [7]. To begin, we randomly selected three out of eight transcripts; two researchers independently read these transcripts and identified a small set of potential codes. These codes were used to develop a mutually agreeable initial codebook to apply holistically to the data. The two researchers then used a copy of the codebook to independently code the three transcripts, while simultaneously refining their own codebook (adding, merging or deleting codes). After this step, the researchers met again to discuss and refine the codebook, and resolve any disagreements on the code assignments. The final codebook contained a two-level hierarchy (12 level-1 codes, 41 level 2- codes), of which the level-1 codes form the high-level themes. This codebook was then used to independently code the remaining five transcripts. For this last step, interrater

agreement between the two coders, measured using Krippendorff's alpha [68], was on average 0.79 (*SD*=0.14, *range*=0.62-1) and the raw agreement 93.8% (*SD*=6.1%, *range*=84.4%-100). Again, the conflicting code assignments were resolved through consensus.

4.3.4 Findings

We detail experience with SoundWatch during the campus walk as well as comments on model accuracy-latency, different system architectures, and the user interface. Quotes are drawn verbatim from the post-trial interview transcripts and in-situ form responses.

Campus walk experience

For the campus walk with SoundWatch, we describe the participants' thoughts on the overall usefulness of the prototype and the variation with contexts. All participants found the watch generally useful in all three contexts (a home-like lounge, office, and outdoors) to help with the everyday activities. 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 [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, participants (8/8) also noticed problems with SoundWatch, the most notable being delay and misclassifications; the latter were higher in outdoor contexts than in others. For example,

> *"Delay might be a problem. When a person came into a room, that person surprised me before the watch notified me [about door-in-use]" (P5)*
>
> *"It doesn't feel refined enough with outside sounds and background noises. The app is perfect for quiet settings such as home and outdoor activities (e.g., hiking in the woods). [While outdoors,] some sounds were misinterpreted, such as cars were recognized as water running" (P3)*

In-situ feedback form ratings reflect these comments, with average usefulness for lounge (4.8/5, *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,

all participants wanted to use SoundWatch for outdoor settings, mentioning that they can use context to supplement the inaccurate feedback (5/8):

> *"Sure there were some errors outdoors, but it tells me sounds are happening that I might need to be aware of, so I can [visually] check my environment..." (P7)*

Besides usefulness, the app usage also changed with location. As expected, all participants chose to enable different sounds for each location; the obvious common choices were fire/smoke alarm, microwave, water running, and speech for lounge; door knock, door-in-use, and speech for office; and bird chirp, car horn, and vehicle running for outdoors, as determined from the system logs. The total number of enabled sounds were also different for each location (avg. 8.3 for lounge, *SD*=1.2; 7.5 for office, *SD*=1.5; and 4.2 for outdoors, *SD*=2.6)—for outdoors specifically, 5/8 participants speculated that the app accuracy may decrease with background noise, and thus deselected all un-important sounds to compensate. For example,

> *"I deselected 'Speech' for outside because I didn't want to know someone was talking outside. It's noisy. [...] I only selected 'car honk', 'vehicle running' and 'siren' [as] they are the bare minimum I need. It seemed to work well then." (P2)*

Model Accuracy-Latency Comparison

Because deep learning-based sound recognition will likely have some error and latency, we asked participants about the maximum tolerable delay and the minimum required accuracy for a future smartwatch app. The most common general 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 type of sound. Specifically, for the urgent sounds (*e.g.*, fire alarms or car horn), participants wanted the minimum possible delay (but would tolerate inaccuracy) to get quick information for any required action. For example,

> *"because I'll at least know something is happening around me and I can use my eyes to look around and see if a car is honking at me..." (P2)*
>
> *"If an important sound is not occurring, I would just be disturbed for a moment, that's all [...] But, if it's an alarm and if this [watch] misses it, that is a problem." (P1)*

In contrast, for non-urgent sounds (*e.g.*, speech, laughing) more accuracy was preferred because 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, it can be really frustrating" (P5)*

Finally, if a sound is a medium priority for the participants (*e.g.*, microwave for P3), participants wanted a balance, that is, a moderate amount of delay is tolerable for moderate accuracy (7/8).

Besides variation with sound type, we asked if the accuracy-latency preference would change with the context of use (home vs. office vs. outdoors). In general, similar to the type of sound preferences, participants erred towards having less delay in more urgent contexts and vice versa. For the home, participants (8/8) wanted high accuracy (more delay is acceptable) because, for example:

> *"I already know most of what is going on around my home. And when I am at home, I am generally more relaxed [so] delay is more acceptable. But, I would not 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 the advantage of having less delay, because *"something may be needing my attention but it's likely not a safety concern" (P8)*. Finally, preferences for outdoors were split: four participants wanted less delay overall with outdoor sounds, but the other four participants did not settle for a single response, saying that the tradeoff would depend on the urgency of the sound outdoors, for example:

> *"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* during the interview.

In general, *watch+phone* was the most preferred architecture for 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 additional advantages of speed and accuracy. This is because in the outdoor context, data privacy was less of a concern for them. For example, P6 said:

> *"Whenever the Internet is available, I prefer cloud for outdoors instead of home/office because of possible data breach at home/office [...] Accuracy problems could be more [outdoors] due to background noise and [thus] I prefer to use cloud if [the] internet is available."*

Watch+cloud was preferred by two participants only for cases where it is hard to carry a phone, such as in a "gym or running outdoors" (P1); others did not share this concern as they reported always carrying the phone—for example*: "I can't really imagine a situation where I would have my watch and not my phone." (P4).* Finally, *watch-only* was not preferred for any situation because of a large battery drain, and a small input area (*e.g.*, for customization).

User Interface Suggestions

Overall, participants appreciated the minimalistic app design, including the information conveyed (identity, loudness, and time) (8/8) and the customization options (mute button, checklist on phone) (7/8). When asked about future improvements, participants suggested three. First, they wanted the app to indicate the urgency of sounds—for example, using vibration patterns or visual colors (*e.g.*, one pattern/color for high priority sounds, and another for low priority sounds). Second, to increase utility, participants suggested to explore showing multiple overlapping sounds (5/8), the most urgent sound (3/8), or the loudest sound (2/8) instead

of the most probable sound as in our design. P4 also said that conveying multiple "possible" sounds could help her 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."*

Finally, two participants (P3, P8) wanted the direction of sound source for outdoor context:

> *"I need to know if the vehicle is running or honking behind me or on the side of me. If it's on the side on the road, then I don't have to do anything. If it's behind me, I will move away." (P8)*

When asked whether they would need direction for home or office as well, they replied no, stating that context awareness is higher for those locations (2/2):

> *"No, not needed for these contexts [home and office]. I know the locations of where the sound [source] could be, if it shows "microwave", it's in the kitchen. If it's "speech", I know where [my spouse] is." (P3)*

4.4 Discussion

Our work in this chapter reaffirms DHH users' needs and user interface preferences for smartwatch-based sound awareness [33,89] 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 tradeoffs, classification architectures, and potential concerns. In this section, we reflect on further implications and limitations of our work.

4.4.1 Utility of smartwatch-based sound classification

How well does a smartwatch-based sound classification tool need to perform to provide value? As both our systems evaluation and user study reveal, this is a complex question that requires further study. While 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? More work is needed here. One way to explore this question would be by releasing SoundWatch—or a similar app—to the public with multiple customization options, then studying actual usage and soliciting feedback. However, this approach also introduces ethical and safety concerns. 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 and of other recent work [33,89] is that users desire *customization* (*e.g.,* which sounds to classify, notification options, sound priorities) and *transparency* (*e.g.,* classification confidence) with sound awareness tools.

4.4.2 Towards improving accuracy

Our user study suggests a need to further improve system accuracy or at least explore other ways to mitigate misclassification costs. One possibility, as our participants suggested, is to explore showing multiple "possible" sounds instead of the most probable sound—just as text autocomplete shows n-best words. Another possibility is to sequentially cascade two models (*e.g.,* see [110]), using the faster model to classify a small set of urgent sounds and to employ the slower model for lower-confidence classifications and less-urgent sounds. End-user customization should also be examined. While installing the app, each user could select the desired sounds and the required accuracies, and the app could dynamically fine-tune the model (*e.g.,* by using a weighted average accuracy metric based on the sound urgency). Finally, as proposed by Bragg *et al.* [13], 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 the DHH user.

4.4.3 Privacy implications

Our participants' showed concern for cloud-based classification architectures: they valued their own "sound" privacy and of others around them. However, uploading and storing data on the cloud has benefits. These datasets can be used for improving the classification model. Indeed, modern sound architectures on IoT devices (*e.g.,* Alexa, 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-on, always-listening sound awareness devices? We see three. First, the users should have control of their sound data. Indeed, P3 corroborated this:

> *"I can see myself potentially using the watch+phone+cloud, if I can, [...] open my laptop and [select/deselect] what [sound data] gets uploaded and who gets to see what. Otherwise I fear that [someone] may misuse something that I don't want them to."*

This data upload can also be customized based on context (*e.g.,* the office might have more private conversations than outdoors). Second, future apps will need clear privacy policies such as GDPR [162] or CCPA [163] that outline how and where the data is stored and what guarantees the users have. Finally, users should always have access to their data and to potentially delete it, in entirety, from the cloud.

4.4.4 Future smartwatch applications

In contrast to past wearable sound awareness solutions [35,49,61], we used commercially available smartwatches, a mainstream popular device that is more socially acceptable than HMDs [35,49] or custom hardware-based [61,102] solutions. A recent survey with 201 DHH participants [25] showed that smartwatch-based sound awareness was preferred over smartphones as well. So, what are other compelling applications of a smartwatch for DHH users? Full speech transcription, a highly preferred feature by DHH users [25,46] is difficult to accommodate on the small watch screen, but future deep learning work could explore highlighting important keywords or summarizing the conversation topics. Sound localization is also highly desired [13,33] and could be explored by coupling the watch with a small external microphone array or designing a custom watch with multiple microphones. But, how best to combine different features (*e.g.,* topic summarization, direction, identity) on the watch is an open question. In another work by our team [33], we investigated different designs for combining sound identity, direction, and loudness, however, this study was formative with a focus on user interface design. Future work should explore the system design of showing multiple features with classification confidence—a challenging problem given the smartwatch's low-resource constraints.

4.4.5 Limitations

Our lab study included a 20-min out-of-lab component intended to help participants think about and experience SoundWatch across "real-world" contexts. While useful as an initial, exploratory study, important pragmatic issues could not be investigated such as user perception of battery life, integration of the watch into daily life, and long-term usage patterns. Future work should perform a deployment study and compare results with our lab findings.

Moreover, our model accuracy results, though performed on real-life recordings of 20 sounds, do not accurately reflect real-world use as other sounds beyond those 20 may also occur. Our tests, however, were enough for our goal to compare the models and contextualize the user study findings. A more accurate experiment would include a *post hoc* analysis of sound data collected from longitudinal watch use.

Finally, we considered our DHH participants (who identify as deaf, Deaf or hard of hearing) as a homogenous group while reporting user study findings. Indeed, past work [13,25] shows that these groups, despite their cultural differences, have synergetic access needs and preferences. Recruiting cross-culturally allowed us to explore solutions for a diversity of users. Nevertheless, future work should examine how preferences may vary with culture and hearing levels.

4.5 Chapter Summary

In this chapter, we reported on 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 that our best classification model performed similar to the state of the art for non-portable devices while 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 contextualize our system experiment results, and also uncover ideas, concerns, and design suggestions for future wearable sound awareness technology.

Chapter 5: Personalizing Sound Awareness

The Chapters 3 and 4 show that sound recognition can provide important information about the environment, human activity, and situational cues to people who are d/Deaf and hard of hearing (DHH). However, the solutions that we built—HomeSound and SoundWatch—or even the commercially available ones—such as on Google Android [164] and Apple iOS [87]—use generic models that are pre-trained on large sound corpora and do not support end-user personalization—such as training on new sound categories (*e.g.,* a new custom home appliance) or a specific sound (*e.g.,* my child's voice or pet's dog bark) [13,25]. As the evaluations of our HomeSound and SoundWatch systems show, personalization is a key desire feature by DHH people.

In this chapter, we present *ProtoSound*, an interactive system that allows users to personalize a sound recognition engine by recording custom sounds (Figure 5.1). Unlike traditional data-intensive machine learning approaches, users can customize a model using only a few sample recordings (*e.g.,* five for each sound). While prior machine learning work (*e.g.,* [116,133]) has performed algorithmic experiments of "*few-shot"* sound recognition, we contribute the first useable system by integrating user-centric features such as: (1) on-the-fly training for difficult-to-produce sounds (*e.g.,* fire alarms, sirens) and (2) handling contextual soundscape variations (*e.g.*, homes *vs.* outdoors). In contrast, traditional few-shot approaches require the full training set to be available beforehand [26,120] and do not generalize well across contexts [16].

Figure 5.1: ProtoSound is a technique to customize a sound recognition model using very few recordings, enabling the model to scale across contextual variations of sound (*e.g.*, water flowing on a stainless steel *vs.* a porcelain sink) and support new user-specific sound classes (*e.g.*, a piano). Images show some example sound categories that were trained and recognized during our field evaluation using an experimental mobile app built off ProtoSound.

To guide ProtoSound's evaluation, we conducted a large-scale survey with 472 DHH participants, which uncovered key personalization preferences such as the minimum number of custom sounds to support and the maximum desired recording effort. We then used these insights to design three experiments: quantitative evaluations on two real-world datasets and a field study. On a dataset of sounds recorded by hearing people in multiple contexts, ProtoSound outperformed the best baseline model by a 9.7% accuracy margin (88.9% vs. 79.2%). The average accuracy (88.9%) was close to the ground truth obtained by manual human labeling (91.3%). On an additional dataset of sounds recorded by DHH people in and around their homes, ProtoSound's average accuracy was 90.4%. In comparison, the dataset's label accuracy rated by a hearing person was 94.5%.

While the above results are promising, they do not reflect an actual system use. Thus, we deployed ProtoSound's end-user training and real-time recognition through a mobile application and conducted a field evaluation with 19 hearing participants—to our knowledge, the first evaluation of few-shot sound recognition in the field. While our ultimate goal is a long-term study with DHH users, demonstrating real-world efficacy and improving the system is an important step before deployments with the target population; hence, we recruited hearing users who could reliably listen to the real-world sounds and evaluate ProtoSound's recognition accuracy. Results show that ProtoSound trained the model on-device through low end-user effort

and accurately learned sounds in a range of acoustic environments (*e.g.*, homes, restaurants, grocery store, parks, and streets). However, errors arose due to recording mistakes (*e.g.*, incorrect labels, overlapping sounds), pointing to a need to develop better user interfaces in the future.

5

5.1 The ProtoSound System

ProtoSound is an interactive system for personalizing a sound recognition model in real-time using few custom recordings. ProtoSound uses prototypical networks [120], one of the most efficient algorithms for few-shot classification and extends the traditional training pipeline to incorporate user-centric features for real-world deployment—such as a technique to accommodate varying contexts of use, and a library of difficult-to-produce sounds preferred by DHH people. Throughout the design of ProtoSound, we worked with individuals of the DHH community, including the lead author who is DHH, and a co-author, who is an ASL interpreter.

5.1.1 System Design

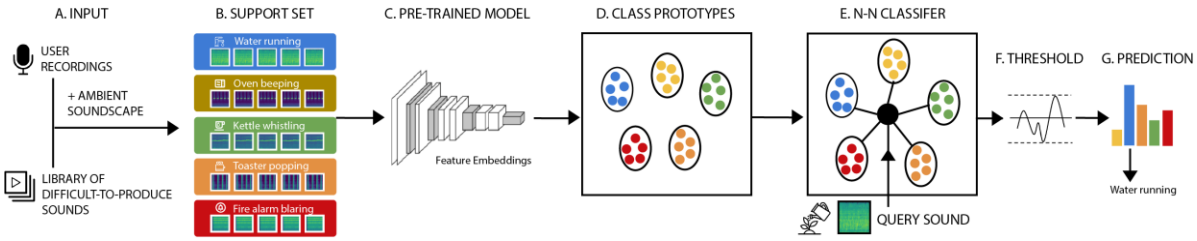


Figure 5.2: Our ProtoSound few-shot sound recognition pipeline. For the desired sound classes, users either record a few samples of sounds on their own or select from our online library of difficult-to-produce sounds (*e.g.*, for fire alarms, sirens) (A). To accommodate contextual shifts, these sound samples are then mixed with the ambient soundspace of a place, leading to the *support set* (B). This support set is then fed into our pre-trained sound recognition model (C) to generate *class prototypes*, which is the representation of each sound class in the feature space (D). During prediction, a query sound is then compared with these stored class prototypes using a nearest neighbor distance metric (*i.e.*, N-N classifier) (E). When our confidence threshold is passed (F), the nearest class is outputted as the prediction (G).

ProtoSound's sound-sensing pipeline involves two phases: model personalization and prediction. Model personalization includes personalizing the model from a set of user recordings (Figure 5.2B). During this phase, *log-mel spectrograms* [41] of user recordings in a context, or samples from our library of difficult-to-produce sounds are fed into the model to extract feature embeddings (Figure 5.2C) We use log-mel spectrogram input features since they have historically shown better performance than other alternatives (*e.g.*,

MFCCs) with CNN architectures [74]. The extracted embeddings for each class are averaged, resulting in class prototypes, which are representations of each class in feature space (Figure 5.2D). These class prototypes are used for predicting a new sound using a nearest-neighbor classifier—that is, we output the class nearest to a query sample by calculating the Euclidean distance in the feature space (Figure 5.2E—G). In addition, to aid real-world use, ProtoSound contains several user-centric features: context generalization, a library of difficult-to-produce sounds, and open-set classification.

Context Generalization

Ideally, the users should record sounds for model personalization in each context. However, in real-life, users may move across auditory contexts (*e.g.,* inside homes to outdoors), and may reuse a model trained in one context in another—for example, a model trained on sounds such as water running in the home could also be used outdoors. Such context shift often introduces novel acoustics conditions (*e.g.,* background noise, changing data distributions) and a model may not generalize well. This is particularly an issue with meta-learning approaches which tend to overfit the model on context specific data [116]. Cross-setting generalization methods increase the robustness of classification algorithms by adapting them to a target context [31].

ProtoSound uses a custom, data-driven cross-setting generalization technique [31]: augmenting the samples procured from a source domain that the model was previously trained on (*e.g.,* a home) with the ambient soundscape of the target domain (*e.g.,* an outdoor location), using the following equation:

$$\text{Target sample} = (1-\alpha) * \text{source sample} + \alpha * (\text{target soundscape} - \text{source soundscape})$$

To determine α, we performed iterative experiments on two benchmark sound datasets (*ESC-50* [103] and *UrbanSound8k* [111]) and selected an optimum value of 0.3. Although we chose a single α value to reduce the number of tunable parameters, it can be set to change based on a particular auditory context shift (*e.g.,* homes-to-outdoors may have a different value than outdoors-to-restaurant). Note that while the soundscape may vary across samples recorded from one context, our context generalization scheme only needs an estimate of the background noise to determine the bounds of a feature space in a context. Beyond accommodating context shifts, cross-setting generalization could also help mitigate other possible acoustic

variations, such as those caused by different recording devices (*e.g.*, a model created on a laptop may be used on a phone).

Existing Library of Difficult-to-Produce Sounds

ProtoSound requires sound recordings for personalization. However, in real-life, there may be sounds that are highly desired by DHH users but do not occur spontaneously for recording (*e.g.*, fire alarms, sirens). To support training for these difficult-to-produce sounds, ProtoSound contains samples of 10 sound categories preferred by DHH people [13,55] (*e.g.*, fire/smoke alarms, babies crying, sirens, bird chirps), procured from a high-quality online library, *FreeSound*. These sound samples were manually cleaned (removing noise, deleting silences) by three hearing authors and are available in ProtoSound's repository. During training, these sounds are augmented with the ambient soundscape of the target domain.

Open Set Classification

Most sound classification tools assume a closed-set classification scenario, with a fixed set of predefined classes to distinguish [2]. In real-world, however, the underlying data distributions of soundscapes are often unknown and can change over time with new classes becoming relevant [2,60]. To accommodate this issue, researchers have introduced open-set classification approaches (*e.g.*, [86]), where an algorithm can also classify a given sound as "unknown". ProtoSound uses the following open-set classification algorithm adapted from Júnior *et al.* [60]:

Let d_1 and d_2 be the respective Euclidean distance of a query sample from the nearest and the second nearest class prototype in feature space. Then, we calculate the ratio:

$$R = \frac{d_1}{d_2}$$

If R is less than or equal to a specified threshold T, the query sample is classified as the same label as the nearest class. Otherwise, it is ignored. Following our internal experiments, we used a T value of 0.6.

In addition to the above algorithm, ProtoSound uses an end-user tunable loudness threshold (default value: 45dB, equivalent to an AC hum). During prediction, any query with average loudness below this value is ignored.

5.1.2 System Implementation

Model Architecture and Pre-training

We implemented ProtoSound using a MobileNetV2 architecture [54]—a state-of-art CNN for mobile devices, measuring about 8MB. Past few-shot learning work (*e.g.,* [55]) did not find improvements from using bigger networks like ResNets [60] due to the risk of overfitting on sparse data [55,59]. We pre-trained the model using a train set compiled from six online sound effect libraries-—*Freesound* [28], *BBC* [152], *Network Sound* [153], *UPC* [154], *TUT* [85] and *TAU* [5]—each of which provide a collection of high-quality, pre-labeled sounds. We selected sound categories for which we found more than 1000 clips, which included a total of 35 common sounds from different contexts (*e.g.,* homes, urban, outdoors, see Table 1). Clips were downloaded, converted to a single format (16KHz, 16-bit, mono), and silences greater than one second were removed, resulting in 38.8 hours of sound data.

We segmented each clip into one second audio segments and computed short-time Fourier Transforms using a 25ms sliding window and 10ms step size, which yielded a 96-length spectrogram covering the frequency range from 20Hz to 8000Hz. We then converted our linear spectrogram into a 64-bin log-scaled Mel spectrogram and generated a 100 X 64 input frame for every one second of audio. We applied Cepstral Mean and Variance Normalization (CMVN) [121] to the log-mel spectrograms before inputting into the model. For training, we used a cross entropy loss function with an Adam optimizer [64].

Selection of the Prototypical Networks Algorithm

We selected prototypical networks as our few-shot algorithm following our performance comparison experiment with five state-of-the-art few-shot learning approaches (MAML [26], FoMAML [26], Reptile [94], ANIL [108] and Prototypical Networks [120]) on benchmark sound datasets (*AudioSet* [30], *ESC-50* [103], and *UrbanSound8k* [111]). Results reflect past work [116,133] with prototypical networks performing the best (*avg. accuracy*=95.6%) followed by ANIL (*avg. accuracy*= 93.4%), Reptile (*avg. accuracy*=91.7%), FoMAML (*avg. accuracy*=90.8%), and MAML (*avg. accuracy*=90.6%). Prototypical Networks also had the lowest training time.

Open-Source Release

For researchers and practitioners to build on our work, a PyTorch-based implementation of ProtoSound with our pre-trained MobileNetV2 model is available at https://github.com/makeabilitylab/ProtoSound. The code can support any number of classes and can run on any device with a Python interpreter. The sound samples can be supplied from live microphone or file input. For live prediction, our code samples the microphone at 16KHz and segments the input into 1-second buffers, which serve as query samples.

5.2 Survey of Personalized Sound Recognition

While past studies have shown that personalized sound recognition is generally desired among DHH people [13,54,55], the specific customization preferences are as yet unknown (*e.g.,* how many custom sound classes are desired in a context, maximum recording effort users are willing to put). We conducted an online survey with DHH participants to better understand these preferences and to shape ProtoSound's evaluation (*e.g.,* the number of classes to use in our experiments).

5.2.1 Participants

We used Google surveys [165], which targets users of the Google Opinion Rewards Android app [166]. Due to our institutional policy, we could not ask about identity (*e.g.,* deaf vs. Deaf) or hearing loss levels. Instead, we relied on a DHH assistive technology screener, and targeted respondents who indicated use of "TDD, TTY, or closed captions" (58% of the selected 472 participants), "Hearing aid" (19%), "Real-time captions (*e.g.,* CART)" (29%), "Android Live Transcribe & Sound Notifications" (18%), and/or "Other hearing assistive devices" (9%) in a survey question. 511 respondents satisfied this criterion, but we excluded 39 who misunderstood our survey (*e.g.,* confused sound events with calendar notifications) or provided invalid responses. The remaining 472 participants were adults (18 and older) across US states and territories with 55% men, 43% women, and 2% of unknown gender. All participants used Android smartphones and were compensated up to $1 USD.

5.2.2 Survey Design

The 10-question survey took about 3 minutes to complete (*avg.*=2 min 47 s, *SD*=3 min 52 s) and asked about the use of the current Android sound recognition feature [164], its usefulness, interest in recording sounds for a future personalized system, and the number of sounds a personalized system should support.

5.2.3 Results

About 34% (162) of the 472 participants used Android sound recognition multiple times a week (22% used it daily). Of these weekly/daily users, 89% rated it useful (40%: extremely useful). Among the remaining 310 participants, the majority (71%) rated its usefulness as neutral and 24% indicated they were not aware of the feature.

Participants were also able to select from a list of options for what prevented them from using sound recognition more frequently. Among the weekly/daily users, 81% were concerned about system accuracy (47%: too many notifications, 17%: incorrectly recognized sounds, 22%: missed sounds, 18%: false alerts) and 33% felt that the recognition was too generic (21%: *"might not recognize some sounds I care about"*, 15%: *"can't select the sounds I want"*), pointing to a need for personalization. Indeed, 73% of the weekly/daily users indicated that they would be interested in recording sounds to personalize the system.

When asked to select the minimal number of sounds a sound recognition technology needs to support in each context (*e.g.*, kitchen, bedroom, restaurant) to be useful, a majority (74%) selected 6 sounds or less (35%: 1-3, 39%: 4-6), indicating that a few medium-to-high priority sounds are desired in a location.

In two open-ended questions, participants specified how much effort (number of sounds and time) they were willing to spend on recording their personal sounds in each context (*e.g.*, kitchen, bedroom). 71% were not willing to record more than 15 sounds and wanted to spend less than 25 minutes for each context.

5.2.4 Discussion

Our findings suggest that nearly a third of our DHH participants use sound recognition multiple times a week and most of those users find it useful (89%). Our results also suggest that ProtoSound could increase usage and value with personalized models, and help users become aware of sounds that are specific to them or their environment. The majority (74%) of our participants indicated that 6 sounds or less could suffice in each context and expressed a desire to not spend significant recording time or effort. These findings suggest that a few medium-to-high priority sounds could cover the needs of a majority of DHH users, and that a low-effort, few-shot experience is important.

5.3 Experiment 1: On Sounds Collected by Hearing Researchers

Our first experiment evaluated ProtoSound on sounds recorded by hearing researchers in real-world settings.

5.3.1 Experimental Setup

Test Dataset: Since commonly used 'synthetic' sound classification benchmarks (*e.g.,* ESC-50 [103], UrbanSound8k [111]) do not mimic the real-world conditions (*e.g.,* background noise, overlapping sounds), we created a 'naturalistic' test set by compiling datasets of real-life sound recordings from two prior HCI works [54,55]. It contains samples for 22 common sounds preferred by DHH people [13,53] and ambient soundscapes recorded by hearing researchers in a total of 21 locations (*e.g.,* homes, university labs, lounges, parks, and urban streets) Thirteen sound classes also exist in the dataset used to pre-train the model; nine are new (Table 1). These recordings were converted to the same format as the train set (16KHz, mono), resulting in 4.5 hours of data.

Table 1: Sound classes in our train (online libraries) and test (real-life recordings) sets. Bolded classes appear in both sets.

Train dataset	Test dataset
Fire/smoke alarm, Alarm clock, Door knock, Typing, Door open/close, Vacuum cleaner, Toothbrush, Toilet flush, Water running, Hair dryer, Wood creak, Sawing, Hammering, Drilling, Dog bark, Cat meow, Cricket, Bird chirp, Engine idling, Vehicle running, Car horn, Footsteps, Breathing, Cough, Snore, Speech, Laugh, Clap, Wind, Train, Helicopter, Aircraft, Gunshot, Glass breaking, Fireworks	**Fire/smoke alarm**, **Alarm clock**, **Door knock**, Doorbell, **Door open/close**, Microwave, Cutlery, Dishwasher, **Water running**, Kettle Whistle, Phone ringing, Washer/dryer, **Dog bark**, **Cat meow**, **Bird Chirp**, Baby crying, **Vehicle running**, **Car horn**, Siren, **Cough**, **Snore**, **Speech**

Tasks: In a meta-learning paradigm, a model successfully *learns to learn* [26] on a set of few-shot tasks sampled from a labelled dataset. Each meta-learning task [26] includes a support set, containing a few examples for model training, and a query set, consisting of examples for accuracy evaluation. In algorithmic terms, a meta-learning task is defined as:

> *Given a support set of N classes (called N-way) and K-samples for each class (called K-shot, where K is small, usually < 10), the aim is to classify samples on a query set along the N classes.*

For our experiments, we used the 5-way, 5-shot setting for two reasons. First, it aligns with past evaluations of generic-model systems [13,54,55] where DHH users found 3-5 medium-to-high priority sounds per context to be sufficient. Second, a low number of classes and samples per class will reduce the user's

recording time—in our survey, 74% participants desired six or fewer classes and 71% wanted to spend less than 25 minutes recording.

Baseline Algorithms: Beyond evaluating our ProtoSound pipeline, we also compared its performance with two baseline approaches: the traditional prototypical networks pipeline [120], which is the current state-of-the-art in few-shot classification [116,133] and a fully supervised method used in commercial systems [55,87,164]. For the supervised method, we pre-trained the model with our train set (Table 1). After pre-training, we replaced the last layer by a randomly initialized linear layer with output dimension of 5 (number of ways) and fine-tuned on the test tasks described in our experiments below.

5.3.2 Specific Experiments and Results

Overall Accuracy

To calculate the overall accuracy, we randomly sampled 100 tasks from our real-life test set, each of batch size 100 containing 25 support samples (5 shot × 5 way) and 75 query samples (15 samples per way). After passing data through the model, we performed a clip-level prediction by aggregating the probabilities for each second of data and outputting the most likely prediction. On average, ProtoSound achieved 88.9% accuracy (*SD*=5.6%), which was significantly higher than the baselines: traditional prototypical networks achieved 79.2% (9.7% less than ProtoSound, pairwise t-test was significant: t_{99}=8.8, p<.001) and supervised fine-tuning achieved 70.6% (18.3% less, t=15.7, p<.001). Improvement over supervised fine-tuning approach is expected since, unlike them, few-shot recognition approaches tend to work well with limited data [116]. Regarding the traditional prototypical networks—the state-of-the-art in few-shot recognition—ProtoSound performed better since it can better handle the soundscape variations (*e.g.*, background noise) in each context, owning to our context generalization scheme. This is better demonstrated in the following context-specific accuracy experiment.

Context-Specific Accuracy

Our test set contains samples from three contexts: homes (kitchen, bedroom, and living room), offices (university labs and lounge), and outdoors (parking lots, parks, and streets). As sound quality may vary across context, we also calculated the context-specific accuracies of ProtoSound and the two baselines. As expected, for all three approaches, the accuracy was higher in quiet environments of homes and offices compared to

outdoors (Figure 5.3a). However, the accuracy difference between quiet and noisy environments (homes vs. outdoors) was much lower for ProtoSound (3.6%) than the baselines (13.8% and 16.0% respectively), suggesting that ProtoSound can better generalize across contexts. Figure 5.4 shows the low-dimensional projections of embeddings obtained from the three approaches in an outdoor context.

Note that we did not calculate per-class accuracies due to the limitation of the few-shot evaluation—each individual test includes a random combination of five classes from our dataset. The accuracy of each class depends heavily on which other four classes are chosen for a specific test (*e.g.*, doorbells perform poorly with phone rings), hence aggregating class performance across multiple tests is counterintuitive.

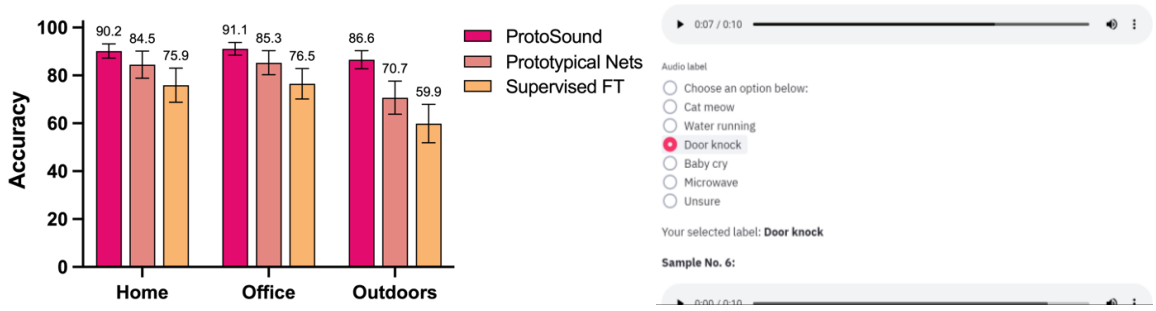


Figure 5.3: (a) Context-specific accuracies of ProtoSound and two state-of-the-art baseline approaches: prototypical networks and simple supervised fine-tuning. Error bars represent 95% confidence intervals. (b) Snapshot of a web-app we built to collect human labels on our real-life test set.

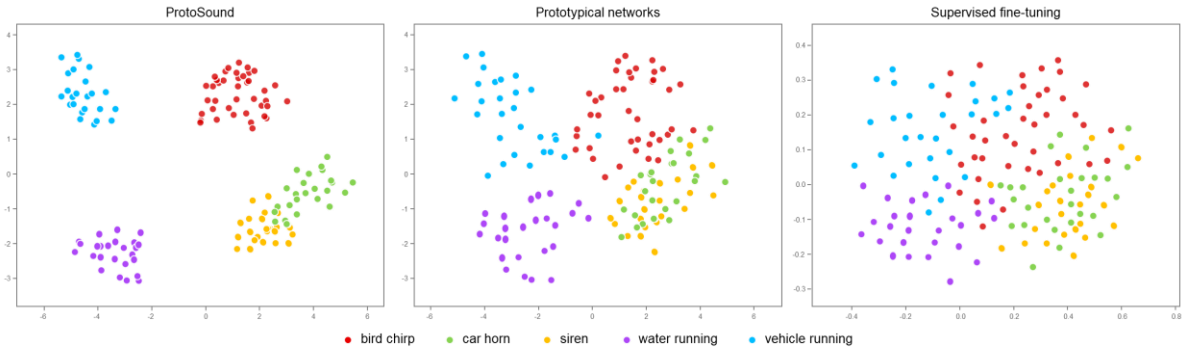


Figure 5.4: A visualization of t-SNE projections of embeddings obtained from ProtoSound and the two baselines: traditional prototypical networks and supervised fine-tuning for samples of five sounds in an outdoor context. Note that even for overlapping sounds "car horn" and "siren", the clusters from ProtoSound are reasonably separated in contrast to the highly overlapping clusters of the two baselines.

Comparison to Manual Labels

To obtain ground truth performance, we recruited people to label our test set. Humans offer an excellent gold standard as they can utilize contextual knowledge from a lifetime of real-world experiences. Similar to

Ubicoustics [72], we created a web-app (Figure 5.3b) that mimicked our ProtoSound model personalization and testing task. The app randomly sampled 5 sounds and 20 samples for each sound from our test set (total 100 samples) which were divided into support (25 samples) and query set (75 samples).

We then recruited six hearing participants from our research group. Each participant listened to the support samples for "learning" and categorized the samples from the query set among the 5 classes (Figure 5.3b). Similar to our open-set classification approach, participants could also select an "unsure" option if they thought a sample did not belong to any of the listed classes. Each participant analyzed three batches, resulting in 6 participants × 3 batches × 75 = 1350 evaluations.

Average accuracy of participants' labels was 91.3% (*SD*=4.8%). Participants revealed two factors that made it challenging to correctly classify some samples: (1) noise (*e.g.*, silence, or too much background noise) and (2) interclass similarities—that is, sounds that were very similar (*e.g.*, doorbells and phone rings). In comparison, our model achieved 92.9% average accuracy for the same setup (*SD*=4.3%), which is close to human performance; a paired t-test was not significant (t_{17}=0.8, p=0.4). On further investigation, we found that, like humans, the errors were most prominent for similar sounding events (*e.g.*, alarm clock and phone ringing), which were often confused.

5.4 Experiment 2: Sounds Collected by DHH People

While the experiment above was necessary to contextualize ProtoSound within prior work, the dataset was collected by hearing people and could lead to representation bias [135]. To combat this, we also evaluated ProtoSound's performance on sounds recorded by DHH participants in a prior study [34].

This data was collected and labelled by 14 DHH participants in locations in and around their home over a one-week period (677 recordings of 243 sound classes, avg. duration=11.5 s). To construct a dataset relevant to our evaluation, we chose participants that had recorded at least 10 classes and at least three recordings per class—resulting in nine participants (P1-P9). The samples were converted to 16Hz mono and silences greater than one second were removed. As this dataset was less balanced than in our experiment above, we could not perform similar granular experiments (*e.g.*, context-specific accuracies).

Class counts per participant, including example classes, are shown in Figure 5.5a. Many of the classes are highly personalized to participants' use cases (*e.g.*, flicking light switch, hearing aid whistle) and indicate that a pre-trained model would not scale well for these individuals. Moreover, existing sound datasets do not contain the requisite samples for several of these classes (*e.g.*, seatbelt alarm) to train a fully supervised model. These characteristics highlight the drawbacks of generic-model systems and reinforce the need for personalization.

ID	Classes	Examples of recorded sounds
P1	14	dishwasher, kettle timer, exhaust fan
P2	23	elevator bell, bedside alarm, dumpster emptied
P3	13	Flicking light switch, kettle, motorcycle running
P4	15	door knock, candle lighter, fit bit alarm
P5	18	oven timer, washer ending, garbage disposal
P6	16	microwave beep, car engine, seatbelt alarm
P7	28	dryer beep, bathroom faucet, oven beep
P8	18	bathwater draining, car running, bathroom door
P9	15	gas stove ticking, hearing aid whistle, doorbell
Total	160	

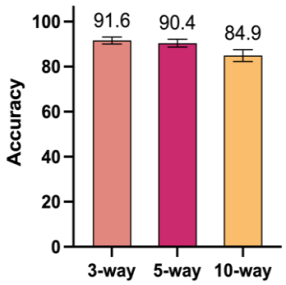


Figure 5.5: (a) DHH participants' recorded sound class counts with examples. Note that many of these classes are highly specific to participants' use cases (*e.g.*, flicking light switch, hearing aid whistle) and thus, require model personalization. (b) ProtoSound's average accuracy for 3-class, 5-class, and 10-class evaluations on DHH participants' recorded sounds.

For our experiment, we evaluated three settings: 3-way (3 classes), 5-way, and 10-way. We trained the model using one randomly selected recording per class for each participant (equivalent to a real-world use case) and used a clip-level prediction. See Figure 5.5b for results. For the 5-way setting—the most desired by DHH people—the overall accuracy was 90.4% (*SD*=4.4%). In comparison, the accuracy of the dataset's labels as rated by a hearing team member was 94.5%. Per-participant accuracies and per-class accuracies for the lowest performing participant (P8) are shown in Figure 5.6. Results were poor for participants P6, P7, and P8 due to two sources of errors: first, similarity among some sound classes led to confusion (*e.g.*, water draining in the bathtub *vs.* in a sink, laundry room fan *vs.* floor fan); second, some recordings did not appear to contain the labeled sounds (*e.g.*, egg cooker, car running for P8). More detailed analysis of user errors can be found in the original work [34].

We also compared performance with a supervised baseline, finding a significant increase in accuracy: for a 5-way setting, the performance difference was 19.7%, pairwise t-test yielded t=16.2, p<.001. Overall, our

analysis showed ProtoSound has the potential to accommodate a wide variety of sounds from our target population.

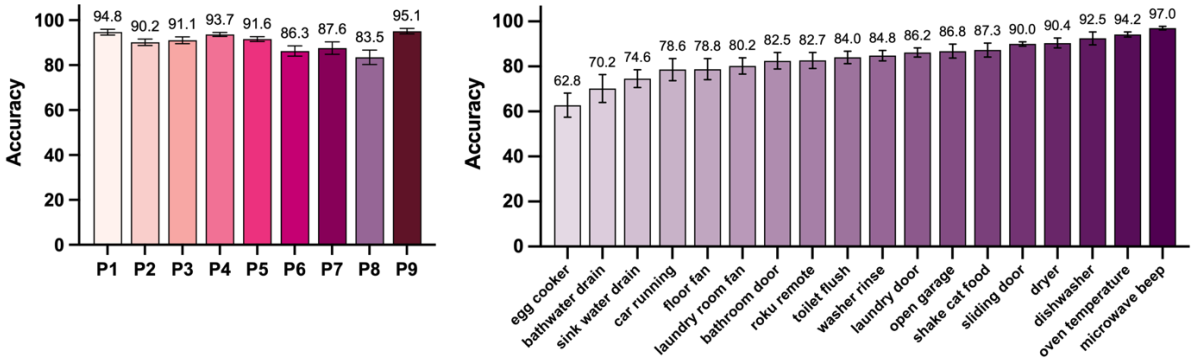


Figure 5.6: (a) Average accuracy per DHH participant for the 5-way setting. (b) Accuracy per-class for the lowest performing case: P8. Note that 'egg cooker" performed poorly due to user recording errors in some samples (missing sound). 'Bathwater drain' and 'sink water drain' performed poorly since they were very similar sounding and were confused with each other.

5.5 Experiment 3: Field Evaluation

Our third evaluation was a field study with 19 hearing participants. While our ultimate goal is a long-term evaluation with DHH users, demonstrating real-world efficacy and improving our pipeline is important before deployments with the target population. Thus, we deployed our ProtoSound technique through an interactive mobile application and recruited hearing participants who were able to evaluate the real-world performance by reliably listening to the sounds and providing feedback on recognition. To our knowledge, our study is the first evaluation of few-shot sound recognition in the field.

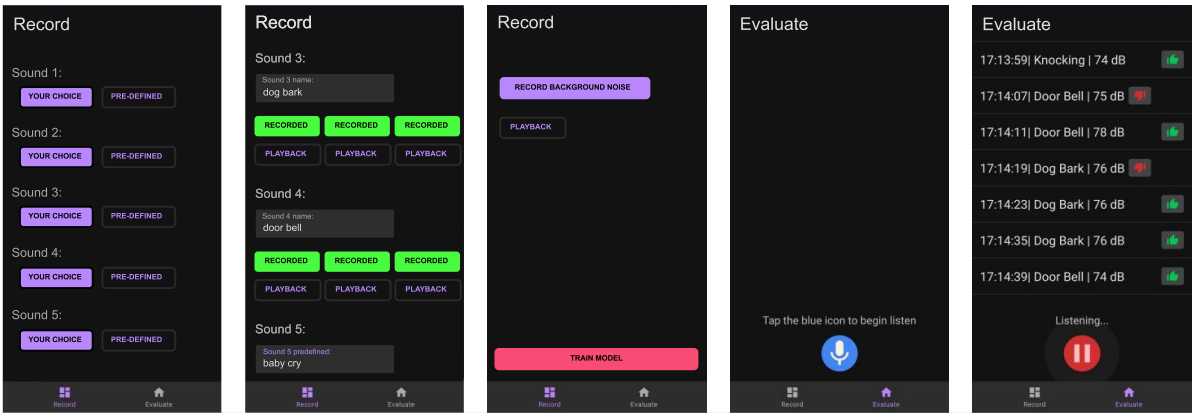
5.5.1 ProtoSound Mobile Application

Our Android-based smartphone app, shown in Figure 5.7, contains an experimental user interface to enable hearing users to train and evaluate a sound recognition model using our ProtoSound technique. We restricted to a 5-way, 5-shot setting (5-classes and 5-samples per class) in each location to reduce our participants' recording time and effort, but ProtoSound can support any setting. Each location uses a separate sound recognition model; for training, users first enter the name of a location (*e.g.,* kitchen, restaurant) and then select the five sound classes for recording in two ways: (1) by entering the name of their own custom sound (*e.g.*, "dog bark", "doorbell" in Figure 5.7b), or (2) selecting a sound from a predefined list (*e.g.,* "baby cry" in Figure 5.7b). For each custom-defined sound, users record the five required sound samples, each of one

second duration. Additionally, they can play back the recording and re-record to correct any errors. For a predefined sound, the app randomly selects five samples from ProtoSound's existing library of sounds (see Section 3.3.2). Finally, users record an ambient soundscape of the location and submit the samples for training (Figure 5.7c).

After training, the app saves the model, which can be used for evaluation at any time by opening the app and clicking on the evaluate tab (Figure 5.7d). For evaluation, the app samples the audio every four seconds (an estimation of average length of all sounds from our test set) and outputs a prediction (Figure 5.7e).

Implementation. To preserve user privacy and support offline use, the mobile app uses a *pyTorch-mobile* implementation of our ProtoSound pipeline and can run fully *on-device*. However, for the study specifically, the app interfaced with a socket.io server located at our institution, and, with each recognized sound, it displayed a binary rating form (correct, incorrect) for users to evaluate the recognition accuracy. These ratings along with other study data (user recordings, system logs) were uploaded to the server for analysis (and deleted after the analysis was complete). The entire app code is open sourced at https://github.com/makeabilitylab/ProtoSound.



(a) (b) (c) (d) (e)

Figure 5.7: User interface of our ProtoSound Android application.

5.5.2 Participants

We initially recruited 20 hearing participants through various social media platforms, but one quit early due to app installation issues. The remaining 19 participants (10 women, 9 men) were on average 38.9 years old

(*SD*=14.0, *range*=21-61), resided in 14 different US cities, and used an Android phone. Participants were compensated $35.

5.5.3 Procedure

The study was conducted remotely due to the COVID-19 pandemic. We emailed step-by-step instructions, with an option to ask for clarifications through email, text message, or online chat. After a link to a short demographic questionnaire, the instructions outlined how to install the app, followed by a short usage tutorial. Then, the participants chose three locations in and outside their home for app evaluation, such that: (1) each location had at-least five "naturally" occurring sounds and (2) at least one location was outdoors (*e.g.*, parking garage, park). Participants completed the recording and the evaluation tasks for five sounds in each chosen location.

For the recording task, participants could choose a predefined sound or record samples for their own sound, although, to ensure that they do not rely heavily on the existing list, they were asked to define at least one custom-defined sound in each location. During recording, the acoustic activity did not have to be naturally occurring; participants could produce the sounds themselves (*e.g.*, by deliberately knocking or turning a faucet on). This ensured high-quality recordings since it may be difficult to get an isolated single class sound sample in real-life settings. To incorporate inter-class disparity, participants were also encouraged to add everyday variations in their recordings when possible (*e.g.*, by recording different kinds of dog barks or vacuuming on different surfaces).

For the evaluation, participants rated 100 app recognitions (by thumbing up or down, see Figure 5.7e) in each location. Contrary to the recording task—where participants were allowed to produce the sounds themselves—the evaluations were performed in “natural” acoustic settings (*e.g.*, during meal preparation for a kitchen location, or during busy weekends in a park). Our criterion was that at least one of the five sounds had to be spontaneously occurring in a location. The app saved the state and displayed a notification when 100 ratings were complete. Participants could then end the test or optionally, rate a few more recognitions. The total evaluation time in a location was about 15 minutes, but it was not necessarily continuous—*e.g.*, participants could evaluate for 5 minutes each during breakfast, lunch, and dinner based on convenience and presence of natural acoustic activity.

After completing all the three locations, participants completed a short questionnaire to provide open-ended feedback on their experience and document examples of any sounds that were consistently correctly/incorrectly recognized or missed altogether during their evaluations.

5.5.4 Findings

We detail our mobile app's usage summary, ProtoSound's overall and context specific accuracy, sources of errors, and comparison to prior approaches.

Usage summary

All participants completed three locations, except P4 who could not evaluate an outdoor location due to quarantine requirements, resulting in a total of 56 locations (5769 sample evaluations). Figure 5.8a shows the locations with some example sounds. In homes, the common locations were kitchen, bedroom, and bathroom. Outdoors, participants selected parks, parking lots, and streets. Other locations included restaurants, cafes, and grocery stores. A total of 171 unique sound classes were recorded.

The total recording time per context (including set-up time, context switch time, and recording time for 25 samples of one second duration each) was on average 10.2 minutes (*SD*=3.6 minutes, *range*=6.1-18.7 minutes). This falls safely below the suggested maximum recording time from our survey (25 minutes), confirming that ProtoSound requires low-effort end-user training. The average model training time (time between submitting samples and obtaining a new model) on the users' phones was 2.4 seconds (*SD*=0.9 seconds, *range*=1.3-4.9 seconds), indicating that ProtoSound supports real-time on-device model personalization. The median total time gap between training and evaluation was 4.0 hours (*IQR*=12.8 hours, *range*=0.1-30.9 hours) and the median total evaluation span was 5.4 hours (*IQR*=18.7 hours, *range*=0.2-51.6 hours). Since the evaluations could be discontinuous, 16/56 evaluations spanned multiple days. This open-ended discontinuous evaluation allowed us to study the real-world applicability of ProtoSound.

Overall and context-specific accuracies

The average accuracy of our app across all locations was 87.4% (*SD*=6.3%). When comparing locations, the accuracy was highest in Bedrooms (*avg.*=92.6%, *SD*=3.8%) and lowest in Restaurants (*avg.*=82.2%,

SD=7.9%), potentially due to differences in noise levels and sound types. Figure 5.8b shows location-specific accuracies.

Note that since participants only rated the sounds that were recognized by our app, our accuracy does not account for false negatives (*i.e.*, any unrecognized sounds). Indeed, in the feedback form, most participants (14/19) self-reported examples of sounds that were sometimes missed by our app with two participants indicating examples of "frequently" missed sounds (keys jingling and bird chirp). At the same time, participants also indicated events that were consistently recognized correctly, such as microwave beeps, door knocks, furniture sliding, door open/close, dog barks, vehicle, cart rolling, and water running, many of which are desired by DHH people [13,55].

Sources of Errors

To determine the sources of errors, we did manual analysis on user recorded samples (through listening, making visualizations), finding that about 10% of the samples contained user errors (this justifies ProtoSound's 87.4% accuracy). Specifically, we found two types of errors. First, a majority of these 10% samples did not contain the labelled sound or contained another sound of interest beyond the labeled sound, thus reducing accuracy. Second, in some cases, samples belonging into different sound categories were too similar (*e.g.*, knocking and chopping, see Figure 5.9) and were understandably confused with each other. This points to the need to develop better user interfaces for recording and annotating sound samples. We return to this point in the Discussion.

Locations	Counts	Examples of recorded sounds
Kitchen	17	water running, chopping, microwave beep
Bathroom	10	toilet flushing, water running, door opening
Park	7	dog barking, footsteps, children swinging
Bedroom	6	door knocks, alarm clock, typing
Parking lot	5	car locking , car door open/close, cart rolling
Street	4	footsteps, vehicle passing, construction
House yard	4	dog barking, ball bouncing, furniture sliding
Café/Restaurant	2	chairs moving, putting cup on table, footsteps
Grocery store	1	cart rolling, sliding door open, stacking carton
Total	56	

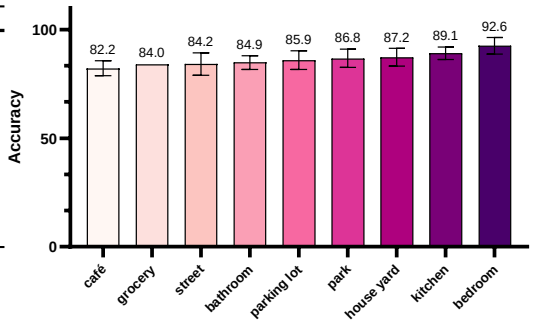


Figure 5.8: (a) Location for our field evaluation with the participant counts and recorded sound examples, and (b) Accuracy of our mobile app in each location. Error bars = 95% confidence intervals (no CI shown for grocery since it only had one count).

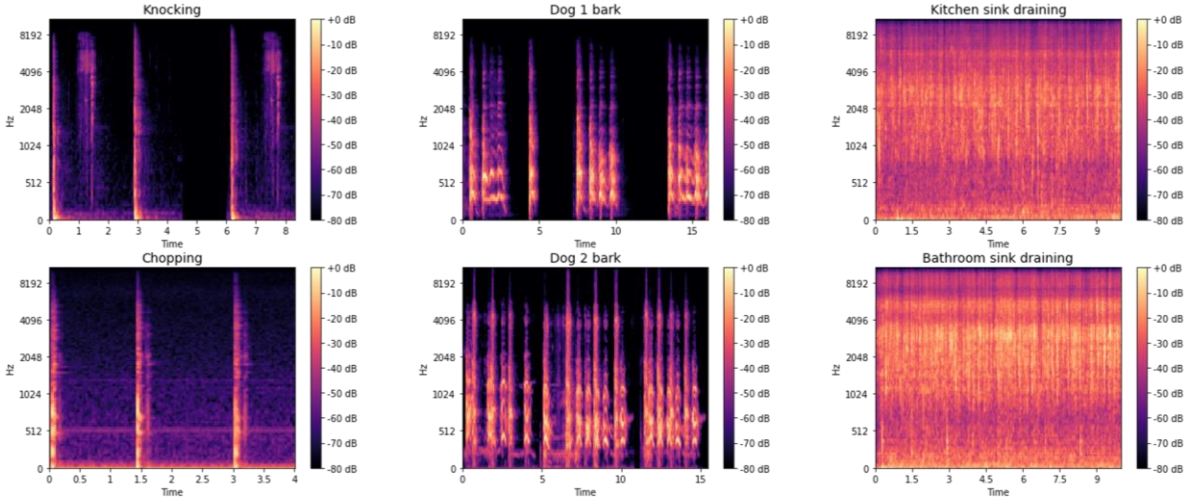


Figure 5.9: Mel-spectrograms of some similar sounding events that were confused with each other: knocking *vs.* chopping, barks of two different pet dogs, and kitchen *vs.* bathroom sink draining. Note the striking similarities in the spectrograms.

Effect of Pre-training and In-Situ Personalization

We also calculated the difference in performance between the sound categories that our model was pre-trained on (*e.g.*, alarm clock, door knock in train set (Table 1), total 2613 samples) and the sounds that were "unseen" by the model (*e.g.*, furniture sliding, children swinging, total 3156 samples). The average accuracy on pre-trained sound categories (91.6%, *SD*=4.7%) was higher than the new sounds (84.0%, *SD*=7.5%), suggesting that pre-training the model with the expected classes that a system may encounter in real-life will increase accuracy

For only the pre-trained sound classes, we also compared performance with state-of-the-art sound recognition systems such as *HomeSound* [54], *SoundWatch* [55], and *Ubicoustics* [72], which use a generic pre-trained VGG16 model [41]. We trained this model on our trainset and evaluated on the field samples, finding that that the accuracy was significantly lower than ProtoSound (71.3%, a difference of 20.3%) a paired t-test yielded t_{29}=9.1, p<.001. This indicates that, while allowing for new, personalized sound classes, in-situ few-shot customization also significantly improves accuracy on existing sound classes by accounting for contextual variations of sounds.

5.6 Discussion

In summary, our work in this chapter makes contributions in the fields of human computer interaction (HCI) and machine learning (ML) fields. Within HCI, we contribute the first personalizable sound recognition system for DHH users. Prior sound recognition systems [54,55,119] use generic pre-trained models, which do not support: (1) sounds unique to a DHH person's use case (*e.g.,* children playing), (2) variations of real-world sounds (*e.g.,* my dog vs. a generic dog), and (3) sounds with insufficient samples in existing sound datasets to train generic models (*e.g.,* footsteps). Our findings show that ProtoSound can accommodate variations in existing classes and support a variety of new, personalized classes in a diversity of contexts through low-effort end-user training.

Compared to prior ML work (*e.g.,* [116,133]) which only performed algorithmic experiments, we contribute the first deployable few-shot sound recognition system by incorporating several user-centric features in the traditional few-shot learning pipeline, such as: (1) on-the-fly training for difficult-to-produce sounds (*e.g.,* fire alarms, sirens) and (2) generalization across contexts. Our experiments show that ProtoSound significantly outperforms the best state-of-the-art few-shot baseline, yielding an accuracy improvement of 9.7% on a dataset of real-life sounds.

We also contribute findings from a large-scale survey (472 DHH users) on personalized sound recognition, insights from the first evaluation of few-shot sound recognition in the field, as well as an open-source plug-and-play system code that can be deployed on any device and a specific Android app implementation. Below we detail further implications of our work and state key limitations.

5.6.1 Graphical User Interface

Recording and annotation interface. Our work focused largely on building a backend pipeline for few-shot sound recognition. However, the user interface is equally important, especially for DHH users who may not be able to verify the contents of their recordings by listening to them—a challenge we observed in samples captured by DHH users. Thus, intuitive sound visualizations are essential for DHH users to better record and label their training samples. Spectrograms and waveform visualizations may be a good place to start; however, these low-dimensional features may not sufficiently represent the sample's quality [34].

Furthermore, our system expects sufficient separation among classes and enough variation among interclass samples. Failing to meet these requirements led to classification errors in our field study, which we expect will increase more when DHH users are unable to recognize auditory similarities among sounds during the training process [34]. Thus, in a parallel project, our team is researching cause-and-effect visualizations (*e.g.*, showing cluster visualization of NN-classifier) with an aim to make end-users really understand how their recorded samples may shape the model [34]. A well-designed user-interface may even further increase our system's accuracy by improving training sample quality and reducing user errors.

Human-in-the Loop applications. ProtoSound is an example of Human-in-the-Loop (HITL) systems, which leverage end-user input in the model building and refining pipeline to improve performance [21,109]. However, user agency in our pipeline is limited to providing training examples. The training process itself is still a black box. Through providing more information about the training pipeline (*e.g.*, by visualizing intermediate model layers, showing measures of uncertainty in model prediction), we hope to support a greater user agency. Our pipeline can also be extended to incorporate reinforcement learning techniques [65] to gradually adapt the model by taking user feedback on the recognition output, for example, using the user interface explored in ListenLearner [136]. Such techniques could capture even greater contextual and temporal variations of real-world acoustic events beyond that is accommodated by limited samples in ProtoSound (*e.g.*, varying cries of a baby or different piano notes). However, while DHH users may be able to validate this output in a familiar location (*e.g.*, in a kitchen) with the help of visual cues, they may find it challenging to do in unpredictable contexts.

5.6.2 Algorithmic Improvements

Acoustic event detection. ProtoSound processes data frame-by-frame using a fixed sampling window. While this worked for our purposes, sound classes vary considerably in length—from short-lived (*e.g.*, a gunshot) to longer events (*e.g.*, thunder)—and selecting an optimal window is challenging. If the window is too small, long-term variations may not be captured. Conversely, if the window is too long, detecting boundaries between consecutive sound events is difficult. Thus, future work should explore acoustic event detection techniques (*e.g.*, sub-frame processing [63] or sequential learning [137]) to automatically segment real-life sound events.

Collaborative learning. Another area of exploration is federated (or collaborative) learning [138], where multiple mobile devices collaboratively learn a shared model while keeping all training data local. This technique is useful for drastically improving model performance without compromising user's data privacy, such as in the Google Keyboard (GBoard) [139] where the gesture typing technique is improved over time by averaging locally personalized models collected from billions of users [167]. Our ProtoSound pipeline is well suited for this task since the generated low-resolution class-prototypes from the model personalization process can be directly uploaded to the cloud, without compromising privacy.

Simultaneous events. While ProtoSound only conveys the most probable sound, our pipeline can also be modified to output multiple simultaneous events. Indeed, in past work [55], DHH users preferred the idea of showing multiple sound events in low confidence situations. However, this could easily lead to information overload, so future systems need to be carefully designed. Similarity detection techniques (*e.g.*, [8]) can be used to group multiple similar events together (*e.g.*, show "appliances" instead of "could be a microwave beep or a dishwasher"). Likewise, systems leveraging contextual information (*e.g.*, location of deployment) can ignore some detected events in a similar way a human understands that a car honk is unlikely to originate from a kitchen.

5.6.3 Socio-Cultural Implications

Deaf culture. While our sound recognition technology is heavily informed by DHH perspectives and past work [54,55], we do not assume it is universally desired or that it will necessarily work, as designed, for all users. Some DHH people may feel negatively towards this technology, especially those who identify as part of the Deaf culture [15,71]. However, our survey aligns with prior work [13,25] and suggests that many DHH individuals find sound recognition valuable, and even more so if it is personalized. Such a tool can be constrained to detect only a small subset of sounds (*e.g.*, a child's cry) to provide essential situational awareness while otherwise avoiding the hearing world. Still, more work is needed with the DHH population to evaluate the accessibility of our system, given diverse preferences and interests.

Privacy. To preserve privacy, our pipeline can run locally on devices without the need to transmit audio data to the cloud. However, uploading data has other benefits such as using it for interactively improving the classification model. Our technique can also support privacy-preserving cloud-based computation since we

compute low-resolution mel-spectrograms of input data, which, while readily identifying speech, make the spoken content challenging to recover.

5.6.4 Limitations

Our work has the following primary limitations.

Five-class setting. First, though ProtoSound can support any number of classes, for two of our three experiments, we used a five-way (five-class) implementation since this most closely resembles what DHH people wanted in past work and in our survey. Specifically, in evaluations of generic-model sound recognition systems [54,55], DHH users enabled only 3-5 medium-to-high priority sound classes in each location to avoid being overwhelmed by notifications. Furthermore, we wanted our field study participants to spend a minimum time recording. ProtoSound's implementation is location-specific—*that is,* users train a separate model for each location, which can be switched manually, or in the future, automatically through a location-aware design (*e.g.*, [55]). Such an implementation can support, for example, 15-25 classes in a home by using a separate model for each room. Nevertheless, while few-shot learning has not yet reached a stage to support more than a few classes [96,124], ProtoSound (and its open-source implementation) can support any setting and we report on performance of different class sizes in our Experiment 2. We also encourage future work to experiment with larger class configurations while using other ways to improve performance (*e.g.*, by constraining to very specific types of sounds, or increasing the number of training samples per class).

Survey recruitment bias. Second, by relying on assistive technology use to identify DHH users (per institute policy), our online survey may have excluded participants who are less likely to use these technologies (*e.g.*, sign language users). We, however, reference a past survey [25] which showed that more than 75% of those who preferred sign language were interested in sound recognition support.

Dataset constraints. Third, we evaluated performance on a real-life dataset complied from two HCI works [54,55] instead of standard machine learning benchmarks (*e.g.,* ESC-50 [103], UrbanSound8k [111]) since these benchmarks use clean sound files and do not mimic many real-world conditions (*e.g.,* background noise, overlapping sounds, context shifts). A notable exception is Google's *AudioSet* [30], but the labelling accuracy of this publicly released dataset is very poor [168]. Nevertheless, we believe we effectively

contextualized ProtoSound's performance by implementing and comparing accuracy with multiple state-of-the-art few-shot baselines on our compiled test set, which contains sound recordings from 21 real-world locations. Future work should collect and extend our experiments with larger, more varied datasets.

Short technical evaluation. Finally, participants in our field study used the app briefly in each location, which while demonstrating promising potential for few-shot sound recognition, does not account for a longitudinal use where a user could be moving through a range of acoustic contexts over time (*e.g.*, home to outdoors to office). While our approach should theoretically handle these contextual shifts, long-term deployments across contexts are needed to quantify the performance over a longer use period.

5.7 Chapter Summary

Sound recognition can provide important environmental, situational, and safety-related cues to people who are d/Deaf or hard of hearing (DHH). Existing sound recognition systems, however, do not support personalization to users' specific desired sounds. In this chapter, we presented the design and evaluation of ProtoSound, an interactive system to personalize a sound recognition engine using only a few custom recordings. ProtoSound was motivated by the prior work with DHH users, the experiences of our DHH authors, and a survey we conducted with 472 DHH participants. Evaluations on two real-life datasets and with an interactive mobile application in the field suggest that ProtoSound can support highly personalized sound categories through low end-user effort, can train the model on-device in real-time, and can handle contextual variations in a variety of real-world contexts. Beyond sound recognition, our personalization technique also has the potential to support applications in other domains such as context-aware assistants, personalized speech recognition, and home automation.

Chapter 6: Exploring Head-Mounted Displays for Conversation Support

While the above chapters of this dissertation focus on non-speech sound awareness, here I report on a more forward-looking work: augmented reality-style head-mounted displays (HMDs) to provide speech conversation support to DHH users. Many DHH people use real-time captioning to access spoken information [92,146]. However, these captions are typically shown on a laptop or large shared screen [92,146]—or, after the recent advent of commercial mobile captioning solutions (*e.g.*, [145]), on a smartphone—which forces the users to shift attention to the captioning screen, drawing their gaze away from the conversational partners and the environment. HMDs has the potential to display captions within the wearer's field of view, which can reduce visual split and increase glanceability [90,99,122].

To demonstrate viability of the HMD-based approach, we designed and performed two evaluations of a working real-time captioning app on Microsoft HoloLens: a long-term autoethnographic study, and a semi-controlled evaluation in a walking scenario. In the first study, I, a hard of hearing individual, used two visual designs of the HMD captioning app during my graduate school classes and group meetings over a six weeks period [46] (Figure 6.1a). Preliminary insights suggest that I could more easily identify speakers and better follow the conversation with the HMD compared to the laptop-based captions. Motivated by this initial finding, we performed a second evaluation where 10 DHH participants used our HoloLens app to converse while walking on a prescribed route on campus [50] (Figure 6.1b). Our findings reaffirmed the potential for HMD-based captions to support glanceable conversations for DHH users, but also identified the need to provide other sound cues (*e.g.*, speaker location, speech tone).

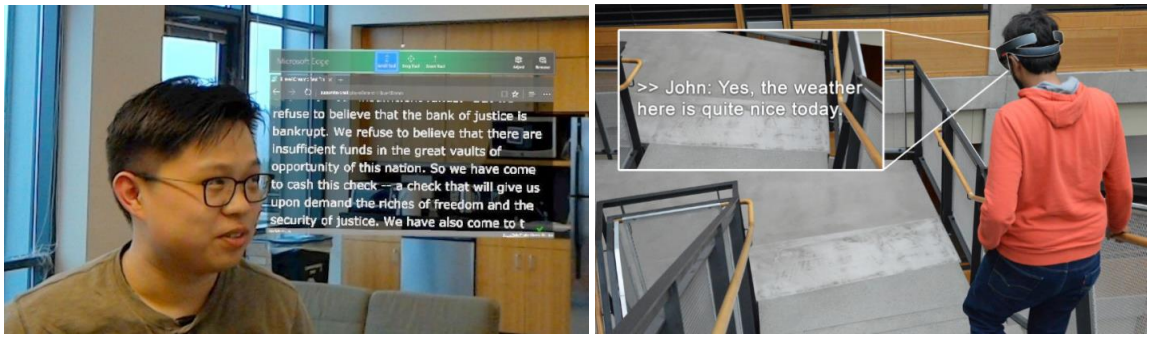


Figure 6.1 My HMD-based captioning app in use during (a) an autoethnographic study, and (b) a semi-controlled evaluation with DHH users.

We then built *HoloSound*, a Augmented Reality system for Microsoft HoloLens that leverages deep-learning and sound sensing to simultaneously provide three key desired sound properties to DHH users in real-time: real-time captioning, sound identity, and sound location [37]. Informed by our past work, HoloSound uses Microsoft Azure API to transcribe speech, our speech recognition model used in HomeSound and SoundWatch work to identify sounds, and a microphone array, ReSpeaker, to localize sounds.

In summary, the primary contributions of this chapter include: (1) empirical results from two evaluations that highlight the potential of HMD to support glanceable conversations for DHH users, (2) an augmented reality-based system that provides three sound cues: speech transcription, sound identity, and source location on an HMD, and (3) key design factors and recommendations for future work.

6

6.1 Study 1: Autoethnographic evaluation of HMD captioning

6.1.1 Initial Prototypes

Informed by prior work [11,130], our own experiences as persons with hearing loss, and building and using our prototypes, we synthesized 13 design considerations for HMD-based AR captioning related to caption rendering (Figure 6.2) and contexts of use (Figure 6.3).

Caption placement: how are the captions positioned in 3D space and do they automatically move (*e.g.,* to track the speaker).

Caption length and size: how many words and lines of text are presented and at what size?

Transcription fidelity: are the words transcribed verbatim or is summarization used (*e.g.,* topic summarization, nouns-only).

Wearer's voice: is the wearer's voice transcribed and, if so, how is it visually represented?

Contextual information: what level of contextual information is supplied about speaker (*e.g.,* speaker names, speaker tone, loudness).

Non-speech information: what non-speech sounds are important and how should they be represented (*e.g.,* dog barking, door opening).

Error handling: how are errors in transcriptions represented and potentially fixed?

Figure 6.2: UI design considerations for AR captioning on an HMD.

Visibility of information: how visible are the captions? *E.g.,* private view (*e.g.,* viewable only by DHH user), or public view (*e.g.,* a large projected display in a classroom).

Conversation group size: how many people are involved in the conversation? *E.g.,* 1:1, medium-sized group (*e.g.,* around a dinner table), or a larger set (*e.g.,* a lecture).

Physical activity: what physical activity are the conversation partners involved in? *E.g.,* all people sitting, main speaker walking while others sitting (*e.g.,* a lecture) or is everyone moving (*e.g.,* walking).

Topic sensitivity and interaction: how may user needs change across conversation topics? *E.g.,* confidential information (*e.g.,* finances), or high emotions (*e.g.,* intimate conversations),

Expected length of interaction: what is the expected length of the conversation?

Relationship with conversational partners: how may user needs change depending on their relationship and familiarity with conversation partners?

Figure 6.3: Context of use considerations for AR captioning on an HMD

To begin exploring this design space, we developed two initial real-time prototypes with the Microsoft HoloLens. Both prototypes use *Streamtext*, a remote online captioning software, to receive real-time captions from a professional transcriptionist. The default caption font used was Arial, size 42, in white color on black background. The font size was often adjusted during the use of Prototype 1.

Prototype 1: AR Windows displays captions in a HoloLens web browser window (Figure 6.4a). Using built-in HoloLens hand gestures, the user can duplicate, resize, and position this window in physical space (*e.g.*, one window above each speaker). Similar to traditional captioning, this prototype can display multiple lines of conversation. Captions scroll up and disappear at the top of the window as new captions are generated.

Prototype 2: AR Subtitles displays one caption window that is placed at a fixed distance in front of the user and moves with the user's head (Figure 6.4b). Similar to video captions, this prototype only shows the most recent generated captions (60 characters). With no option to resize or position the captions, AR Subtitles provides less user control than AR Windows.

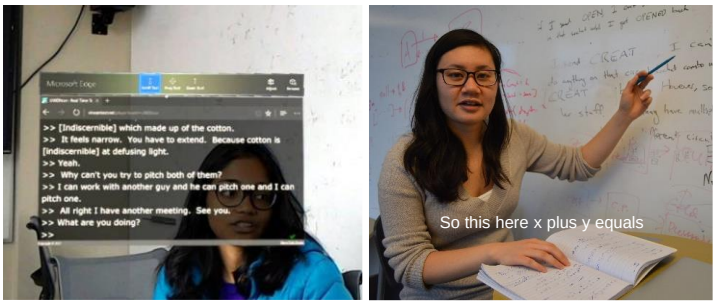


Figure 6.4: (a) Prototype 1*: AR Windows* displays captions in a HoloLens web browser window. Captions can be overlaid close to speakers or visual materials, such as lecture slides in 3D space. (b) Prototype 2: *AR Subtitles* displays one caption window that is placed at a fixed distance in front of the user and moves with user's head.

Figure 6.5: Illustrative figures showing: (a) 1:1 meeting with AR Subtitles, (b) group meeting with AR Windows, (c) our lead author positioning a caption window close to a speaker with AR Windows.

6.1.2 Evaluation

To gain preliminary insight into the benefits and limitations of AR captioning on HMDs and to help inform the design of our future user studies, we conducted an autoethnography. Autoethnography is a qualitative research method which includes a reflexive and analytic account of personal experiences, and connects those experiences to wider social and cultural groups [22]. I adopted the role of researcher-participant and documented my experience with our prototypes over a 45-day period. In the subsequent texts, I refer to myself in third person voice, "Jain", for objectivity and formalism.

Method

Jain is a 26-year old graduate student with severe-to-profound deafness. He uses bi-lateral hearing aids in both ears, can speak and speechread well, and is not fluent in sign language. He relies on real-time captioning for classes and meetings. Jain used the two prototypes in 10 instances of his regular group meetings and classroom lectures. The total usage time was about seven hours. AR Windows was used in six instances, three group meetings and three lectures, while AR Subtitles was used in four group meetings. Jain also used

laptop-based captions for about 30 instances during the same academic quarter, most of which occurred before we built the HMD prototypes.

Jain documented his experiences of each HMD session in the same day using a semi-structured approach, describing the context of use, his emotions, events that stood out, and his general experience. His notes contain a total of 7,053 words. We used a thematic approach [14] to analyze the data based on both inductive (*e.g.*, "hardware limitations") and deductive (*e.g.*, "glanceability") themes. The first author led this analysis, guided by multiple discussions with other team members as analysis iteration occurred.

Findings

In terms of overall preference, Jain was initially split between the laptop and HMD captions, especially due to the discomfort of wearing the HoloLens (see below). However, by the fourth session, he preferred the HMD to the laptop. Here, we highlight differences between the two prototypes and between the HMD and laptop approaches. Quotes draw directly from Jain's notes.

Glanceability. Jain felt he could switch more quickly between attending to captions and attending to the speaker when using the HMD than with laptop. For example, after the fourth session (3 hours of total use), he wrote:

> *"HoloLens was better than laptop since I could see [both] [speaker's] lips and, the captions…". Consequently, Jain could "make some more face-to-face contact with people [speakers] using HoloLens [as compared to laptop captions]."*

Using HoloLens, captions were also closer to visual materials such as lecture slides, which increased the ease of access to information, particularly in cases where the speaker pointed to visual aids. For example,

> *"[While using AR Windows in a lecture,] I could see his hands pointing [at] various math equations on the screen, as he said 'derivative of this [pointing at slides] leads to this.', which is hard to follow with captions on a laptop."*

Caption Display. With both HMD prototypes, Jain attempted to overlay captions on or below the speaker's face by using hand gestures to position captions in AR Windows or by positioning his head in AR Subtitles. For example,

> *"I positioned my head so that captions appeared to come out from his mouth. I also at times, positioned [my head so that] the captions [were] below his face, center aligned with his face… so I could see both the current captions and the speaker's lips in one focus view."*

Additionally, for AR Windows, Jain tried to place one caption window for each speaker in 3D space to reduce visual dispersion (see Figure 6.6b) but used only one window for speakers who were close to each other. When looking at lecture slides, Jain positioned the captions directly below the slides to avoid visually obstructing the slide material (Figure 6.6c).

Comparing our two prototypes, Jain preferred AR Windows during group meetings with multiple, seated speakers because he could set up a caption window for each speaker. He noted that *"though the captions were not always in my view [like AR Subtitles], I was able to speechread speakers while seeing their captions* with [close to] them." Instead, with AR Subtitles, *"captions felt disconnected from speakers as [captions] appeared at a fixed distance from me."* For multiple moving speakers, however, he preferred AR Subtitles as captions remained in view while the speakers moved. Instead, for AR Windows he *"had to move the screen [caption display] around."* He speculated that he would prefer AR Windows overall if *"captions could automatically move with the speakers in future."*

For all cases involving a single speaker (*e.g.,* 1:1 meeting, non-interactive lecture), Jain preferred AR Subtitles to AR Windows as he could position himself in a way that captions were close to the speaker, and he was also *"able to read captions when I moved my head"* (*e.g., "for taking notes",* see Figure 6.5a).

Social Aspects. Jain reported feeling unusually noticeable and socially awkward wearing the HoloLens, both because it is unusual to see people wearing the HoloLens and because AR Windows required conspicuous physical gestures to configure captions. For example, after a four-person group meeting with AR Subtitles, he noted: *"I felt people were little distracted with my HoloLens, especially [name withheld], who hasn't seen*

me with it before. She got used to it eventually but in the beginning, she appeared a little uneasy." And after a lecture with AR Windows,

> *"I think the hand gestures movements […] distracted students. [The instructor] deduced what was going on in HoloLens. But he told me [after the lecture, that], if he didn't know, he would be like 'What are you doing in my class? Playing a game while I teach?"*

One conversation partner commented that the HoloLens partially obstructed Jain's face, making it difficult to have a natural face-to-face interaction.

Hardware and UI Limitations. As the current version of HoloLens is heavy and conducts heat, Jain could only sustain 42 mins of continuous use on average. This usage time was shorter when more head movement was required (*e.g.*, to follow a moving speaker). For example, after a brainstorming session, Jain *noted "a [slight] pain on my neck which got worse over those 30 mins"*. As well, the HoloLens display screen is not fully transparent, which impacted Jain's ability to see in a dimly lit room. For example, after a group meeting in a dim room, Jain noted that *"…it was hard to speechread with the HoloLens display obstructing the view…"* Jain also switched to laptop captions after 15 mins of a 1:1 meeting because the HoloLens made it difficult to read the other person's shared computer screen.

We also note some UI problems with AR Subtitles, where a single caption line was placed at a fixed distance from the wearer's head and moved horizontally and vertically with the head. If the speaker was at a much greater distance from the wearer than the captions were, glanceability was reduced: "*it was hard to focus on both speaker and captions."* Conversely, speakers sometimes occluded the captions if they came too close to the wearer. Jain felt that these issues would be addressed if *"the [caption] window could automatically align [with speakers] in the depth space."* Finally, a single line of captioning was often not enough to understand well (*e.g.*, when captioning errors occurred), and in some cases Jain used laptop captions in parallel so he could review the caption history.

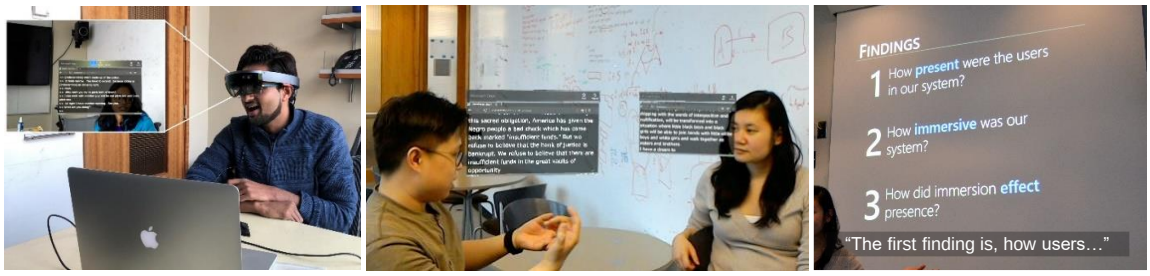


Figure 6.6. Examples to highlight some of our findings: (a) AR Windows in a 1:1 meeting session with caption window overlaid on the speaker; (b) AR Windows in a group meeting with one caption window close to each speaker; (c) AR Subtitles caption display below a lecture slide.

6.1.3 Study 1 Summary and Discussion

Our findings from this preliminary autoethnographic evaluation suggest that a real-time HMD-based captioning approach is promising, particularly in increasing visual contact with speakers and access to other visual material (*e.g.,* lecture slides). At the same time, we identified several improvements that could be made to our designs, such as auto-tracking speakers and providing less obtrusive control over the interface (*e.g.,* to resize and move captions). The HoloLens allowed us to quickly prototype our 3D caption interfaces, but an ideal form factor would be lighter, smaller, and would not obscure the wearer's eyes.

Future work should also investigate the social implications of HMD-based captioning compared to traditional approaches. For example, with an HMD, the captions are only visible to the wearer, which offers privacy and security advantages over a non-wearable display. However, this lack of visibility may also impact how others respond to the device, since the social acceptability of an HMD can be impacted by whether others perceive it to be assistive [106]. Moreover, captions on a private versus shared display may affect conversational dynamics since hearing conversation partners cannot see captioning lag or errors.

Ultimately, real-time HMD-based captions have the potential to reduce visual dispersion, increase user agency, and support a wide range of use scenarios (*e.g.,* mobile) for its users. We used autoethnography with the early prototypes to gain initial insight into HMD-based captioning and to inform further user studies. In the following sections, we investigate our further plans, which include refining the prototypes and conducting larger user studies to generalize our findings beyond a specific user.

6.2 Study 2: Identifying Needs in Mobile Contexts

While the above work examined HMD captioning in stationary contexts (*e.g.*, group meetings, lectures), this technology could be especially useful while conversing in *mobile* contexts (*e.g.*, while walking, in transit), since balancing among looking at multiple sources of information (*e.g.*, the speaker, the environment, captioning) is common in such contexts and could drastically increase the visual split for DHH users. Furthermore, captioning solutions are typically designed and studied in stationary contexts and are not conducive to use in situations when a person is moving. Some research has explored captioning on smartphones [79,134,169] for mobile contexts, but this approach still requires shifting attentional focus away from the speaker and environment and onto a handheld device.

While HMD captioning shows promise for mobile contexts, before designing and evaluating an HMD captioning solution (or any technological solution), it is vital to outline the needs and requirements of DHH people in mobile contexts. While prior work has investigated the communication needs and potential solutions of DHH people in general [19,20,38], this work has not focus on mobile contexts which could present new challenges such as varying background noise, changing lighting conditions, and increased visual attention split.

To identify needs and potential technology solutions, we conducted a formative study with 12 DHH participants.

6.2.1 Participants

Twelve volunteers (five males, six females, and one not disclosed) were recruited through email, social media and snowball sampling (Table 6.1). Participants were on average 34.5 years old (SD=15.3, range 18–66). Eight had profound hearing loss, while the remaining four had at least mild hearing loss. Most reported onset as congenital (N=9). Ten participants used a hearing device: two used cochlear implants, seven used digital hearing aids, and one reported using both. Nine participants (excluding P5, P6, and P12) employed speechreading at least some of the time.

For communication, eight participants preferred sign language, and four preferred to speak verbally. Three participants reported that when conversing verbally with hearing people, they understood more than 81% of the speech; three understood 61-80%; four understood 41-60%; and the remaining two could barely understand speech (<20%). Participants received $25 as compensation.

Table 6.1: Demographics of participants in Study 1 (P1, etc.) and Study 2 (R1, etc.), covering age, gender, hearing loss, lip-reading, preferred communication method, and percentage of speech understood in verbal conversations. "ND" means Not Disclosed.

ID	Age	Gender	Hearing Loss	Lipreads?	Prefers	% Speech understood
P1, R1	23	F	Profound	Yes	Sign	41-60%
P2, R5	18	ND	Profound	Yes	Verbal	> 81%
P3, R3	24	F	Profound	Yes	Oral	41-60%
P4	55	M	Severe	Yes	Sign	41-60%
P5	32	F	Profound	No	Sign	< 20%
P6	21	F	Mild	No	Oral	>81%
P7	28	M	Profound	Yes	Sign	41-60%
P8	35	F	Profound	Yes	Sign	61-80%
P9	66	M	Profound	Yes	Sign	> 81%
P10, R4	32	M	Severe	Yes	Oral	61-80%
P11, R2	23	F	Moderate	Yes	Sign	61-80%
P12, R6	57	M	Profound	No	Sign	< 20%
R7	28	F	Profound	No	Sign	< 20%
R8	31	F	Profound	Yes	Sign	41-60%
R9	28	F	Profound	Yes	Sign	41-60%
R10	54	F	Profound	Yes	Sign	21-40%

6.2.2 Method

The study procedure included a two-part, semi-structured formative interview and took about one hour. We investigated challenges and strategies for conversation in moving contexts, ideas for future technologies, and, for those showing preference for HMDs, a brief captioning technology mockup on HoloLens. Participants communicated with the research team by typing in a shared Google Doc. When desired, sign language was also used for minor clarifications or small talk.

Part 1. The session began with a questionnaire to collect demographics and background on the participant's hearing loss. The researcher then conducted a semi-structured interview on the frequency, location, and social context of mobile conversations, problems encountered during these conversations, how the participants handled those problems, and the impact of physical space (*e.g.*, architectural layout). Three mobile scenarios were explicitly explored: walking, public or personal transport, and other recreational activities such as sports, hiking, or kayaking.

Part 2. We presented three real-time captioning ideas for mobile contexts (phone, smartwatch and HMD; see Figure 6.7) and asked for the participants' preferences with rationale. If a participant preferred the HMD for at least one scenario, we introduced the experience of HMD-based captions using a physical mockup on HoloLens. Because we wanted feedback on the general concept of HMD captioning, we asked participants to imagine a future lightweight version of the HMD for this exercise. Our HoloLens mockup presented a single line of scrolling text from an example script at a fixed distance from the wearer. Participants briefly wore the HoloLens and explored our mockup while walking around the room and turning their heads at different angles. We asked for their thoughts about the overall concept and to describe or draw any other design ideas to support mobile conversations.

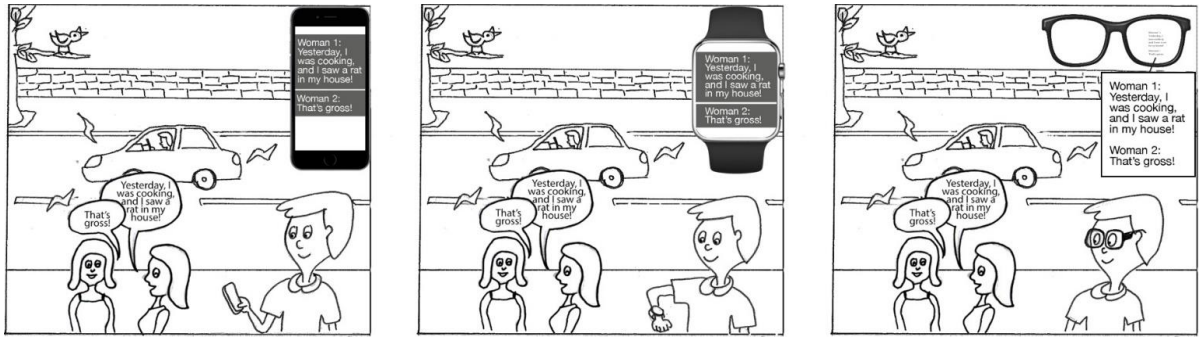


Figure 6.7: Images shown to the participants in Study 2, to guide them in choosing their preference among three potential real-time captioning devices: phone, smartwatch and HMD.

Data Analysis

We thematically analyzed [14] the interview transcripts that had been recorded in the shared Google Docs. One researcher first scanned the transcripts and identified 962 excerpts to be coded across all participants (M=80.2 excerpts/person, SD=19.0). Next, the first researcher and a second one defined three deductive codes based on prior research [19,20] (*adaptive strategies*, *maladaptive strategies*, *form factor comparison*).

Both researchers then categorized the excerpts using the deductive codes and an affinity diagraming process to allow for inductive themes [10]. Through this process, we identified six inductive codes (*social*, *environmental*, and *personal challenges*, *technology use during moving conversations*, *characteristics of moving conversations*, and *user requirements*). Finally, a third researcher reviewed the codes assigned by the first and second researcher, agreeing in 93.6% of cases. Disagreements were resolved through consensus.

6.2.3 Findings

We report on findings related to moving conversations, technology use, social and environmental challenges, maladaptive and adaptive communication strategies, comparison of HMD, phone and smartwatch for captions, and design suggestions for future technologies.

Characteristics of Moving Conversations

Participants mentioned having conversations (spoken or via sign language) while walking to or from meetings, classes, and social activities (N=11), as well as on public transport (N=11), in cars (N=9), and during other recreational activities, such as hiking, playing pool, and kayaking (N=6). All participants had engaged in moving conversations in the preceding week, although the typical frequency varied from only once (P7) to 20 times (P12) a week (M=8.1, SD=5.1).

For conversational partners, most participants reported having moving conversations with both deaf and hearing classmates (N=11), followed by coworkers (N=6), family (N=4), and strangers (N=4). Further, P2, P3, and P4 stated that they never engage in mobile conversations with their hearing bosses or professors since, as P4 explains: *"Those conversations [with boss] are critical. We [sit and] always face each other when speaking."*

Technology Use in Moving Conversations

To communicate with hearing people during moving conversations, eight participants reported using a messaging or note taking app on their smartphone. Most (N=7) resort to using an app only when it is too difficult to hear the other person (N=5) or when the environment is noisy (N=2):

> *"I don't usually use technology other than hearing aids in moving conversations. I will occasionally use my phone to type something if it's impossible to hear." (P10)*

Participants use these apps on public transport (*N*=7), when riding in cars (*N*=4), and while walking (*N*=2). For example, *"BIG [a note taking app] makes it easier to communicate in the dark while driving"* (P6). Five participants reported that they rarely use technology in moving conversations because as P12 explained, *"I don't usually write [type on phone] and move at the same time because it is too challenging"*.

No participants reported using a real-time captioning system during moving conversations, but three participants had used transcription services while moving, including *Microsoft Translator* (P12), *Google Speech to Text* (P4), and voicemail transcriptions (to read rather than listen to voicemail, P8). P12, for example, discussed using a speech-to-text app:

> *"The Microsoft translator [speech-to-text app on phone] makes it easy to have a walking conversation because I can hold my phone in whatever orientation I need to see my surroundings […]. [Further, using the group chat option] each person can talk/read on their own phone to have the conversation." (P12)*

P12 also explained some problems with this app:

> *"The MS Translator isn't perfect because it demands that I split my attention and [also] have one [hand] holding the phone."*

Social Challenges

Our second key finding relates to social challenges experienced during moving conversations.

Moving conversations are superficial. Participants explained that moving conversations lack social connection and convey limited information compared to stationary conversations (*N*=9). Because of split attentional demands, moving conversations were described as being brief and shallow, consisting mainly of small talk (P3, P7), comments and jokes (P4), questions and answers (P11, P1), and "*syncing up information*" (P8). Consequently, the potential for social connection is compromised. For example, P1 reported feeling socially isolated, "*It's hard to make the [walking] conversation smooth enough to go deep… I feel like left out all the time*." Further, four participants mentioned that it is difficult to convey and receive complete information from a moving conversation. For example,

"[While driving,] often we are relying on smartphones, GPS, transportation, and many other distractions. [So,] conversations will go on strange tangents because my focus is divided." (P8)

"It's really hard to walk and talk and lip read... The experience overall often can be negative because 90% of the time you're unsure if you conveyed/understood the conversation well." (P7)

Contexts and communication method. Participants faced challenges with speechreading (*N*=5) or signing (*N*=3) based on the context. For example, because they have to focus on looking ahead, P1 and P10 face difficulty with speechreading in the car when they are driving, and P1 and P7 could not speechread while hiking or biking. Signed conversations were difficult mainly during driving because of the need to keep hands on the wheel (P5, P8, P12). In addition, P5 and P12 explained that understanding signs requires the listener's complete visual attention, which is difficult to provide as a driver, especially when the passenger is not fluent in signing.

Lack of participation from others. Seven participants said that hearing people do not understand and accommodate communication needs in moving conversations. Group conversations with hearing people are especially difficult since they do not stop talking when the deaf person needs to look away (*N*=5). As P10 explains,

"If I need to look away for some reason, a deaf person will automatically stop talking and resume when I'm ready. A spoken conversation doesn't have that type of natural stop and start…"

Challenges during recreational activities. Recreational activities can be especially challenging (*N*=4) since they often require instruction and feedback during complex movements (*e.g.,* martial arts, yoga), greater physical exertion (*e.g.,* running, soccer), or chaotic environments (*e.g.,* white-water rafting). For example, P11 is sometimes unable to hear her yoga instructor calling out the next move and P12 reported missing instructions because he could not see his coach's signs during wrestling. These two participants also explained another challenge: their instructors found it difficult to demonstrate a movement and concurrently explain through signing. Pl2 said:

"[In] martial arts: you have an instructor showing how to move the arms, hands, body, etc. while talking to describe it. Well if they have to "talk" by signing, then how the [heck] do they also show you how to hold your arms in the proper position?"

Environmental Challenges

Participants also noted challenges related to the environment such as building layout, noise, and variable lighting.

Visual attention split. Seven participants felt that moving conversations were more challenging than stationary ones due to the need to divide attention between the conversation and environment (*e.g.,* cars, people, traffic lights, obstacles). This attention split is especially challenging on uneven pavement or hiking trails (*N*=7). For example, P5 said: *"When the sidewalk is bumpy, I have to focus more on looking ahead."* Nine participants were concerned that attention split during conversations posed a safety issue, as P8 explains, "*Safety and communication often compete for my attention in a walking conversation.*" Four participants were particularly concerned about safety while driving.

Space concerns. The design and composition of indoor space affected moving conversations. Large spaces and/or spaces with predictable layouts made it easier to move and face the other person (*e.g.,* grocery stores or museums, as mentioned by P8). In contrast, narrow spaces that required the speaker to be in front or behind the participant reduced understandability (*N*=3). Conversations in cars were also difficult when the speaker was not facing the participant (*N*=7). Another concern was the distance from the speaker; three participants reported not being able to hear speakers who were far away. There was also a social element to interpersonal distance, such as when P6 commented on the difficulty of conversing with strangers: "*it might be awkward to stand closer to them to hear their voice.*"

Background noise. Moving conversations were impacted by background noise from announcements, bus engines, and other passengers' movement in public transit (*N*=4); rain and wind while walking and hiking (*N*=3); and other traffic while driving (*N*=2). Comparing walking and buses, P2 said:

"I do find conversations on buses easier than walking on a busy street. While there is background noise, it is consistent."

P1 and P2 both found it easier to converse at night because *"environments tend to be quieter and less busy"* (P1).

Lighting concerns. In contrast to P1 and P2's comments above, nine participants felt it was easier to have moving conversations during the day when they could more easily see people's faces to lip read or see sign language. Indoor lighting could also negatively affect speechreading when it was too bright (P5, P11) or too dark (P8, P11).

Accommodation Strategies

To address the challenges noted above, participants employed a variety of adaptive and maladaptive strategies.

Adaptive. Verbal adaptive strategies included asking the speaker to repeat (*N*=3), repeating what was heard back to the speaker (P2), asking the other person to speak louder (P3), explaining that they are hard of hearing (P6), and controlling the flow and length of conversations (*N*=2). The most common non-verbal adaptive strategies were adjusting one's seating position to be next to or across from the speaker on public and personal transport (*N*=7), walking side-by-side (*N*=4), turning to face the speaker (*N*=3), using mirrors to see speakers in the backseat of the car (*N*=2), and choosing a quiet path to the destination (*N*=2). P9 and P10 mentioned that they would sometimes stop moving to aid comprehension and participation:

> *"I stop everybody walking and resume the conversation until a point has been addressed, and then start walking again" (P10).*

Maladaptive. Maladaptive strategies included avoiding group conversations (P6) or talking to strangers (*N*=6), postponing a conversation (*N*=6), avoiding spoken conversations altogether (*N*=6), and briefly pausing a conversation (*N*=2). For example, P3 does not talk to Uber or Lyft drivers to avoid "*spending the energy on* it." P11 and P12 chose parts of a conversation to pay attention to rather than the whole conversation. Some participants (*N*=7) avoid conversations based on context. For example, P12 prefers to focus on the scenery while hiking than on conversations, P10 defers important conversations when walking, and P8 lets a phone call go to voicemail when walking.

Envisioning a Real Time Captioning System

We next discuss the social implications of a captioning system, and compare phone, smartwatch and HMD devices.

Social implications. After viewing the mockup in Figure 6.7, all participants said they would use real-time captioning in at least one moving conversation scenario (walking, transit or other recreational activity). However, seven participants wanted to employ captioning selectively because of how the captions may affect conversation quality. As P11 explains,

> *"I always prefer direct communication with hearing people. If technology or interpreters are involved, there is always a distance between me and the other person. It diminishes the quality of the human connection."*

Four participants were concerned that their communication practices and skills would change because of a captioning system. P1 wondered if a system would change the use of sign language, while P9 worried about a potential loss of communication skills by relying too much on a system. P8 wondered if *"a system would change my lifelong pre-disposition to look at people's faces when they talk*." P10 preferred to choose when to follow a conversation, which may be difficult if he had to *"[look] at captions all the time."*

Comparing devices. Eleven participants (except P5) preferred the HMD in at least one moving context (walking, transit, or recreational) (Table 6.2). The perceived main advantage of the HMD was that it would reduce attention split by positioning captions within the user's gaze (N=6). Four participants also felt that the HMD could improve both the quality of a moving conversation and social connections:

> *"[The HMD] would help me be more "connected" to people [compared to smartwatch or smartphone] since I can look at the people [while reading captions]." (P11)*

P9 wanted to use the HMD to overhear group conversations to increase a sense of social inclusion:

> *"I want to be unobtrusive, whether I'm part of the conversation or eavesdropping on the conversation (just like hearing people)."*

For high-contact sports, four participants preferred to use smartphone since they were concerned that the HMD "*would be knocked off easily"* (P11). These four also preferred to use smartphone rather than an HMD on a public or personal transportation because of the option to sit and focus on the display, a common behavior in these contexts. Only one participant preferred a smartwatch (for transit). The most common perceived disadvantage was that the small watch display would only accommodate a limited number of words (*N*=8), which would affect readability (*N*=5) and the ability to follow a conversation (*N*=3).

Participants also mentioned disadvantages of the HMD (*N*=4) and phone (*N*=7). Three participants were concerned that the HMD would be too visually distracting and overwhelming. P6 explained that the HMD might give her motion sickness, a recurring condition for her. Concerns for the phone included display size (*N*=5), needing to look down at one's hand (*N*=5) and the effortful process of holding up one's hand, especially during long conversations (*N*=4).

Table 6.2: Participants' preference (if any) among three devices for captions (HMD, phone, watch) in different moving contexts.

	Walking	Transit (bus, car)	Recreational
HMD	All but P5	P1,P2,P4,P8,P9,P12	P2,P4,P7,P9,P12
Phone	P5	P3,P5,P10,P11	P3,P5,P10,P11
Watch		P6	

HoloLens for captioning. While we used the HoloLens to help ground discussion about HMD-based captioning and asked participants to envision a more futuristic, streamlined device, participants who tried HoloLens (*N*=11) expressed concerns about heaviness and bulkiness (*N*=8), comfort (*N=6*), and how it may hinder mobility (*N*=2). Six participants mentioned social acceptability concerns due to the large form-factor; all suggested using glasses or contact lenses instead to display captions. All eleven participants, however, appreciated HoloLens as a good prototyping tool for evaluating future HMD captioning devices: *"I know future devices would be smaller and [would] fit on my face better, but HoloLens works good for testing"* (P11).

Design suggestions for HMD captioning. When asked about future improvements for HMD captions, participants wanted the ability to go back through a conversation history if they missed something (*N*=3), adjust settings and caption attributes, such as position, contrast, and font (*N*=2), and convey additional information, including: who was speaking (*N*=3), the position or distance of the speaker (*N*=3), and noises in the environment (*e.g.,* door opening) (*N*=2). Five participants wanted captions to remain on top of the speaker regardless of the user's head movements since the caption movement was *"dizzy[ing]"* (P6) or *"distracting"* (P11).

Technology design sketches. When describing their ideas for future captioning technology, four participants also sketched designs. Three of them (P2, P4, P9) extended conventional glasses to include a small screen for displaying captions to avoid wearing the heavy HoloLens. P5 sketched an integrated GPS and voice-to-text system for a car (Figure 6.8a); she was concerned about the technology burden posed on deaf people and wanted to avoid wearing a new technology while driving. P2 sketched an integrated display for contact lenses and proposed a wrist-worn display (Figure 6.8b). Finally, to share the technology burden with hearing people, P9 sketched a t-shirt design that displayed captions that could be read if the speaker is wearing that t-shirt.

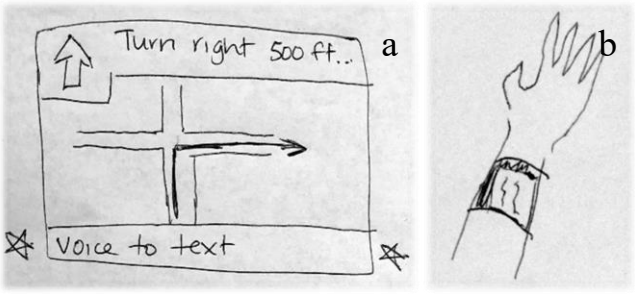


Figure 6.8: Future technology design sketches by participants. (a) P5 proposed integrating speech-to-text system with the car GPS to reduce information overload from multiple devices. (b) P2 proposed a wrist-worn screen to display captions. They said this would be more readable than a smartwatch and more portable than a smartphone or HMD.

6.2.4 Study 2 Summary and Discussion

Prior work has explored communication challenges for DHH people, such as background noise [20], inability to hear a speaker without visual contact [38], and visual attentional demands in signing conversations [70]. Our findings indicate that these challenges also manifest in moving contexts though with greater severity. Mobile contexts impose greater attentional demands than stationary contexts, due to greater visual attention split and context changes, which include topographical, spatial, and noise variation. We also uncovered new

challenges; mobile conversations for DHH people are generally brief and shallow, greatly affected by environmental and spatial characteristics, and have limited technology support and use.

Of particular significance is the context of recreational activities such as yoga, dance, or wrestling, which require that individuals receive time-critical information. As P12 explained, verbal instruction for physical movements are most effectively conveyed simultaneously with a demonstration of the movement. If instruction and demonstration are sequential rather than parallel, the meaning is *"diluted."* For non-signing users who are focused on the activity, paying attention to vocal information provided by instructors is difficult. For signing users, their hands might be otherwise occupied.

Due to attentional and mobility demands, mobile context technologies need to be carefully designed. Captioning technologies, in particular, need to be portable, should adapt to changing contexts, and potentially employ automatic speech recognition. As is mentioned in past work [79,134] and also corroborated in our findings, phones and smartwatches are not typically preferred for mobile contexts since these devices demand split visual attention, a dedicated hand to carry, or are too small for displaying captions. HMDs have the potential to reduce this attention split but need to be lightweight, comfortable, and unobtrusive for broad acceptance. As our participants suggest, other form-factors like glasses and contacts could be leveraged for captions in the future. Displaying captions on non-wearable artifacts, such as in-car GPS systems (P5), would further reduce visual dispersion and the need to carry a personal device.

6.3 Study 3: Evaluating HMD captioning in Mobile Contexts

Informed by the above study findings, we built a proof-of-concept HMD captioning prototype and evaluated it in a real-life walking scenario with 10 DHH users. Our primary goals were to assess whether the use of HMD-based captions increased conversation accessibility and decreased attention split for walking conversations. We chose walking because walking and public transport were the most common moving conversation scenarios mentioned in our previous study and, compared to transit, walking requires more consistent visual attention.

6.3.1 Prototype

Our new HMD-captioning prototype improves on our earlier prototype used for autoethnographic evaluation. To inform the design of this new prototype, we had two aims: (1) increase conversation accessibility and (2) reduce the visual attention split between the environment and captions. For accessibility, we chose to always place the captions in the user's gaze—even when the wearer is not facing the speaker(s). To reduce visual split, the captions were automatically projected onto surfaces (*e.g.*, walls, floor) using the HoloLens' environment mapping capability if desired by the user.

Users could configure the number of lines, length of each line, the font size, and the distance of captions from the eyes (2m, 4m, 8m, or projected onto surfaces; see Figure 6.9). To reduce jitter by head movement, captions stayed at the same location until the user's gaze exceeded 25 degrees. We used Streamtext [170], a remote online captioning software, to receive real-time captions from an on-site professional transcriptionist; however, future versions could incorporate ASR engines. Captions were rendered in white, Arial font.

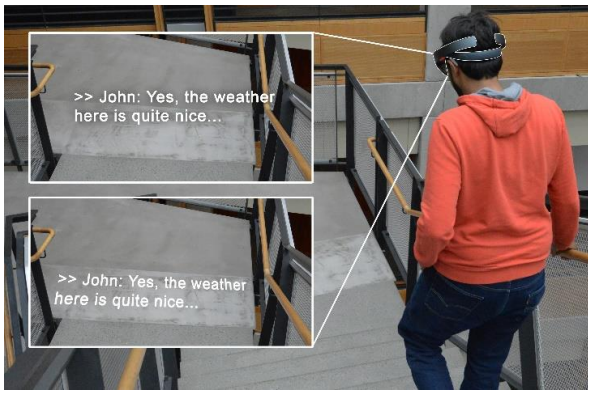


Figure 6.9: An image showing two different options for caption placement in our HoloLens prototype: (a) at a fixed distance from the eyes (here, 4m), or (b) automatically projected onto a surface (here, a floor).

To prepare for Study 3, I myself evaluated our prototype in a campus building via three walking sessions (*avg.* 44mins each). While moving, I interacted with a total of eight people and evaluated different caption configurations. I found that I could understand speakers better using our prototype; however, wearing the heavy HoloLens device for long periods proved tiring. Thus, we decided to limit Study 3 sessions to 20mins. For the captions, I preferred two lines of text, 60 characters per line, an angular font size of approximately 0.75 degree, and surface projection. These configurations were Study 3 defaults, but users could adjust them if desired.

6.3.2 Participants

We recruited 10 participants, including six participants who preferred the HMD idea from Study 2 and four new participants recruited through snowball sampling (Table 6.1). The four new participants (R7 to R10)

were on average 35.2 years old (*SD*=10.9, range 28–54). All were female with profound hearing loss. R7 and R8 developed hearing loss at 2 years of age while R9 and R10 had congenital hearing loss. R9 and R10 used a hearing device (hearing aid). All four participants preferred sign language for communication, and three participants (except R7) employed speech-reading. While conversing verbally with hearing people, R7 could barely understand speech (<20%), R8 and R9 could understand 41-60%, and R10 21-40%. Finally, of all 10 participants, three (R3, R5, R8) used real-time captioning in daily life and the remaining used sign-language interpreters. Participants received $25 as compensation.

6.3.3 Method

The study procedure took on average 57 minutes (*SD*=10.3) and included a walking scenario (*M*=24 mins, *SD*=5.4) and an open-ended interview (*M*=30 mins, *SD*=12.4). The study began with participants briefly viewing our HoloLens prototype. A researcher adjusted the font size, number of lines of text, and the length of each line if the participants desired. After a quick test, the participants walked in a university building wearing our prototype. Two researchers accompanied the participant. Researcher 1 walked alongside the participant and initiated a conversation on casual topics, such as family, weather, food, or events in the city. Researcher 1 also wore a wireless microphone, which relayed the conversation to a remote professional transcriptionist to generate real-time captions. Researcher 2 observed the interaction and recorded notes. Participants could interact with other passersby if they so desired.

After the trial, we conducted an open-ended interview about the experience and solicited feedback about the prototype. Both researchers also asked follow-up questions based on Researcher 2's observations. For this interview component, participants communicated with the researchers by typing in a Google Doc (*N*=6) or verbally (*N*=4). Similar to Study 2, we also used the captions from a transcriptionist and sign language to facilitate communication. Finally, participants completed a short multiple-choice questionnaire asking about how much they depended on captions during the trial and how their speech understanding in daily life compared to that when using our prototype.

Data Analysis

We retained the professional transcripts and used them to conduct an iterative coding process on the interview responses [14]. One researcher scanned the responses, developed an initial codebook, then iteratively applied

the codes to the transcripts, updating the codebook as necessary. The codes were applied to transcripts as a whole and not individual excerpts. The final codebook contained eight codes (*e.g., caption placement, impact of the environment*). A second researcher then independently assigned codes to each transcript using the final codebook. Krippendorff's alpha across all codes was on average 0.65 (*SD*=0.24). Conflicting assignments were resolved through consensus.

6.3.4 Findings

We first describe overall reaction to the prototype, followed by specific themes that arose.

Overall reaction. When asked about the overall experience, all participants mentioned using our prototype to understand at least some part of the conversation while walking. For example, R6, who cannot comprehend oral speech, said:

> *"Being deaf, I can't have that [walking] conversation with a person without some assistance. I need another person who can sign, or another captioning device (Like MS Translator on the phone). Both of these would entail looking away from the speaker. So HoloLens might be better in that regard. […] With the phone translation, I also need to dedicate a hand to holding the phone."*

On a scale of 1 (mostly unintelligible) to 5 (mostly intelligible), we asked participants to rate how well they typically understand their everyday walking conversations (sign language or oral). Participants reported an average of 2.4 (*SD*=1.17). When asked how well they understood conversation during the study session with the HMD, participants reported an average of 3.8 (*SD*=0.79) A pairwise t-test was significant; $t(18) = 3.13$, $p=.006$.

Four participants (R2, R4, R6, R7) appreciated the ability to follow the conversations while looking ahead (Figure 6.10a). R4 explained, *"The big thing with this is you can look where you want and still follow along with the conversation."* However, R3, R4, R5 and R8 found captions to be occasionally distracting. For example, R5 said:

> *"When I was trying to formulate my own responses, I would find the captions quite distracting and, in cases like that, I wish […] that I could look away from [the captions], at my discretion."*

Caption preferences. Most participants (except R3, R5) preferred the default configuration for the captions (0.75-degree font, two lines, and 60 characters per line). R1 said:

> *"Two lines were good... I can look away from the captions for a little bit and then something that somebody said will still be there when I look back."*

R3 and R5 increased angular font size to 1 degree, and R5 used three text lines. Regardless of preference, all participants appreciated the customizability of the captions.

Apart from captions, participants wanted the HMD to display speaker identification cues (*e.g.*, name, location) (N=5) and environmental sounds (N=3). R6 wanted an indicator that someone was talking based on which he could *"decide to listen or not"*. He also thought it would be useful to display voice tone and volume:

> *"Tone of voice (and volume) is a big one, for example indicating volume by font size and tone by font. You could also have something else visually indicating it (e.g. sometimes tv/movies put a little music symbol up for music)."*

To reduce information overload, R9 proposed that we:

> *"include everything [additional sound cues] and maybe have options for people to filter out because I can imagine some people going crazy with all the information overload."*

Visual attention split. To understand the conversation, all participants used both speechreading and captions. The post-trial questionnaire showed average dependence on captions of 3.2 (SD=1.03) on a scale of 1 ("*I did not look at captions*") to 5 ("*I only looked at captions*").

Participants who prefer to converse orally (R3, R4 and R5; Table 6.1) reported looking at speakers more than captions (average dependence on captions: 2.3; SD=0.58; Figure 6.10c). They used captions only to fill in missed parts of the speech or confirm their understanding of the speech. For example, while involved in a group conversation with three passersby, R3 and R4 missed speech during speaker transition and used captions. The remaining participants, who preferred sign language for communication, focused on captions

more than speakers (average dependence on captions: 3.6; *SD*=0.98). For example, *"[I was] mostly focused on captions [and did] not really [look at] faces"* (R6).

When not actively engaged in a conversation, participants alternated between looking at their surroundings and captions. R4, R6 and R8 responded that they focused more on their environment, while R2 focused more on captions to see if somebody was speaking. For the other six participants, we could not ascertain a clear indication of attention split between captions and environment from their responses. But based on the post-trial questionnaire, all six indicated that they looked at their surroundings much less than in a typical moving conversation in their everyday lives.

Caption placement. Six participants (all but R4, R5, R7, R10) appreciated our idea of captions staying in the user's field of view. For example,

> *"I liked that the captions are still going even when I am not looking at you in case you are talking to me when I am not looking." (R2)*

However, five of those six participants wanted to be able to turn-off captions. For example, R3 noted:

> *"[When] I'm cooking something and sometimes I don't want to see captions, so turn[ing] [captions] on and off would be a good option... it would give you a break from seeing them all the time."*

In contrast, R9 commented:

> *"[I] would just leave it on all the time because I can easily ignore [the captions] if I don't want to pay attention."*

R3 also wanted the option to *"move captions down here [below the display] or on the side [periphery]"* to focus more on the surroundings. R6 wanted to move the captions closer to the speakers in his field of view: *"I can see [the benefit of] moving the captions a few degrees towards the speaker."*

The remaining four participants (R4, R5, R7, R10) wanted the captions to be positioned above speakers (like speech bubbles) so the only way they could see captions is if they were looking at speakers. R4 also mentioned another advantage of speech bubbles:

> *"It would allow for the captions to be like a whole paragraph, so somebody could speak, I could look away, but they would have like a whole backlog of things, so I could follow along."*

However, when asked how they would notice speakers outside their field of view (*e.g.*, behind them), R4 and R7 wanted to be able to *"toggle between the two"* caption positioning options, *i.e.*, moving with the user's head or speech bubbles (R4).

Figure 6.10: Images illustrating findings from Study 3. (a) R7 followed the conversation while looking ahead. (b) R8 held the railing to guide her movement; it was difficult for her to walk on stairs wearing the HoloLens. (c) R4, who prefers to converse orally, looked at the speaker instead of captions.

Impact of the environment. Six participants found it harder to walk on stairs since they had to split their visual attention between the captions and the stairs (Figure 6.10b). For example,

> *"When I walked on stairs I needed to look at the stairs. I had to also pay attention on captions. So I was a little nervous that I might step wrong." (R1)*

R2, however, reported that after an initial period, *"I was more aware of how [the HoloLens] works, so I adapted a little bit."* Additionally, five participants who walked in broad daylight had trouble looking at the captions. For example, R3 said: *"the HoloLens display is not bright enough to accommodate for natural lighting."*

HoloLens device. As in the preceding study, participants reported that the HoloLens is heavy (*N*=7), has a limited field of view (*N*=4), is not fully transparent (*N*=3), and draws attention (*N*=3). Three additional insights related to the device emerged. First, as mentioned, our captions moved only when the angular deviation of the user's head exceeded 25 degrees. R4 found this stabilizing technique *"laggy,"* but R5 liked it and described it as a *"cool stabilizing technique."* Second, as a person walks, the HoloLens device learns and adjusts to the new environment, which made the captions appear jittery for a moment. R5, noticing this, said: *"it [captions] kind of flickered a little bit when I was staring at architecture that [are] like rounded columns."* Finally, R5 commented that the automatic depth adjustment of captions in 3D space was a *"really cool feature."* However, R4, R7 and R8 wanted the captions placed at a fixed distance *"coz otherwise my eyes would have to constantly adjust [to different depths]"* (R7).

6.3.5 Study 3 Summary and Discussion

In this study, we evaluated a proof-of-concept HMD captioning prototype on Microsoft HoloLens with 10 DHH participants in a semi-controlled walking scenario. Our findings demonstrate the promise of always-available captions rendered on an HMD and also help identify important areas for future work such as the incorporation of non-speech sounds (*e.g.*, speaker location, environmental sounds). We now detail design implications, future work, and study limitations.

Design Implications for HMD Captioning

Based on our findings, we propose the following design recommendations for HMD-based captions, which can be investigated and validated in future work:

Text alignment. Captions should automatically align close to the speaker or background to reduce the visual attention split between captions and the environment.

Adapt to changing context. Caption color and background should automatically change based on lighting conditions.

Wearer's voice. The HMD should have an option to disable the wearer's voice to minimize information on the screen.

Contextual information. Besides captions, HMDs should convey information such as speaker name and location, speech tone, and environmental cues (*e.g.*, door opening).

User customizability. HMDs should allow customization of caption position, contrast, font, and background. Prior studies also support the need for customizability [80,81].

Future Work

Our initial findings show that HMD-based captions can support communication access in mobile contexts. While the use of HMD-based captions seems to improve the attentional balance between the speaker(s) and navigating the environment, future work should explore this in depth. One potential solution may be to limit the amount of real-time text by displaying keywords or a summary of text; however, extracting this text automatically remains a difficult challenge. Another potential solution may be to position captions directly above speakers and use visualizations to identify speakers outside the field of view (*e.g.*, see [49]). As in [46], users should be able to control the placement of captions and to temporarily or permanently silence them as necessary.

Future work should also explore the role of tactile feedback as a complementary information channel. While tactile information is lower bandwidth than visual information, tactile information often imposes a lower attentional demand [23] and may be useful for some tasks (*e.g.*, notifications, indicating direction of sound).

Finally, some assistive technology efforts have been criticized in the deaf community as "manifestations of audist beliefs" (*e.g.*, [24]), where the technology burden is imposed on deaf people to accommodate hearing communication standards. We acknowledge this important concern. Indeed, an example from our data that attempts to counteract this imbalance is P9's idea that all conversation participants (deaf and hearing) should display real-time captions on their shirts. Captions are predominantly visual, and therefore potentially

provide functional equivalence to deaf and hearing users who are comfortable reading captioned broadcast and online programs daily. We emphasize that our studies and technology designs were informed by our own experiences as DHH individuals, our previous work with DHH participants and wearable sound awareness technologies [46,49], and perspectives drawn from the literature [78,99]; however, we are a team composed of technologists. We also emphasize the diversity among deaf viewers–some prefer captioned videos, and others prefer signed videos. Future work should continue to engage with the DHH community to ensure that we are asking the right questions and pursing appropriate solutions.

Study Limitations

Our study has four primary limitations. First, our findings on attention split relied on self-report. Future work should conduct a comparative gaze tracking study with and without HMD captions to more accurately determine how users' visual attention shifts in moving conversations. Second, we used lapel microphones and a professional transcriptionist for the real-time captions. Future work should explore tradeoffs in transcription quality, lag, and the impact on the conversational experience in mobile contexts that may come with automated captions. Third, we evaluated while walking for a short period in a single building. Future work should consider longitudinal deployments in a variety of contexts. Finally, we largely relied on a live two-way Google Doc to communicate with participants during interviews (similar to [49]); however, some participants may have been more fluent in sign language than English (we did not collect data on this).

6.4 Study 4: Supplementing Captioning with Other Sound Cues

The above studies confirm the promise of HMDs to provide glanceable, always available, and private speech transcription to d/Deaf and hard of hearing (DHH) users in both stationary and mobile contexts. At the same time, our study participants also identified the potential usefulness of conveying other aspects of sounds, such as the identity of non-speech sounds (*e.g.*, door open, someone clapping) [54] and the location of the sound source [33] on an HMD.

Past work has investigated showing these sound attributes, albeit individually—for example, by using external microphone arrays to localize sounds and show them on an HMD [35,49]. While conveying these attributes individually was deemed useful [49,50,55], in real life, these information sources often co-exist,

and must be conveyed simultaneously to users. Indeed, in a recent large-scale survey, DHH people expressed a strong desire for receiving non-speech sound cues along with the speech transcription [13].

We investigated an early prototype of an augmented reality (AR) based system, called *HoloSound*, that leverages advances in deep learning and sound sensing to provide three key desired sound properties to DHH users simultaneously in real-time: speech transcription, sound identity, and source location (Figure 6.11). HoloSound uses a speech-to-text API to generate a transcription that can be positioned in 3D space, a deep-learning engine to display the three most recent non-speech sounds (*e.g.*, doorbell, knocking), and an external microphone array to visualize the direction of at most four sound sources in the vicinity.

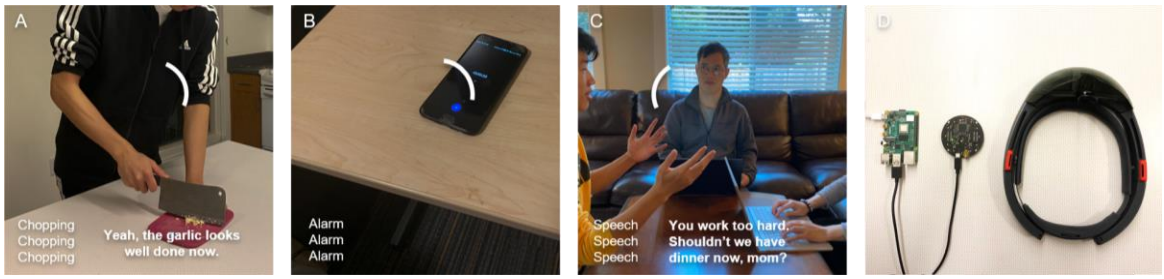


Figure 6.11: Illustrations of HoloSound showing sound identity, source location, and speech transcription. The three most recent sounds are shown at the bottom left of the display, the locations of at most four simultaneous sound sources are shown as circular arcs in the center, and the speech transcription is either shown as subtitles or can be positioned close to the speakers in the 3D space (not shown). For more details, see video.

6.4.1 The HoloSound System

The HoloSound system consists of two parts: an app running on the HoloLens and an external microphone array, ReSpeaker [171], retrofitted on top of the HoloLens (Figure 6.12). A cloud-based server is used to interface the portable microphone array with the HoloLens and to run the sound recognition engine. Figure 6.11 shows the preliminary user interface. Below, we detail the three key components of HoloSound, which are also demonstrated in the supplementary video. The system is open sourced on GitHub: https://git.io/JJaHe; see a demo video.

Figure 6.12: The three components of HoloSound system.

6.4.2 Speech Transcription

Many DHH users use speech-transcription [32,69], and a recent survey [25] showed that HMDs were the most preferred wearable device for speech feedback. To transcribe speech, HoloSound uses Microsoft Azure's *Speech-to-text* API [172]. To accommodate multiple contexts-of-use, we offer two views for displaying the transcribed text: *windows* and *subtitles*, both informed from prior work [46], which used a human transcriptionist rather than automated methods to generate captions for display on a HoloLens device.

In the windows view, the goal is to reduce the visual split between the transcribed text and the speaker, hence the user can place the text windows on top of the speakers using the HoloLens' pinch gesture (Figure 6.13a). This view could be more suitable when the speakers are stationary (*e.g.*, in a group meeting [99]). We use the HoloLens' 3D spatial mapping feature to recognize the environment and automatically position the captions at an appropriate depth near the user's desired spatial location.

In the subtitles view, a single text block appears at a fixed distance in front of the user and moves with the user's head (Figure 6.13b). This view is analogous to video captioning and could be preferred when the speakers are moving (*e.g.*, in a lecture setting [46] or while walking [50]).

For both views, the text scrolls up and disappears as the new transcription is appended. Users can also customize the transcribed text's font size (default: 0.75° angular), the number of lines (default: 2), the width of each line (default: 60 characters), and the distance of captions from the eye (2m, 4m, 8m, or the default: projected onto background surfaces (*e.g.*, walls) [50]). To stabilize jitter, the text block stays at the same location and smoothly drifts along when the user's head moves by at least 25°.

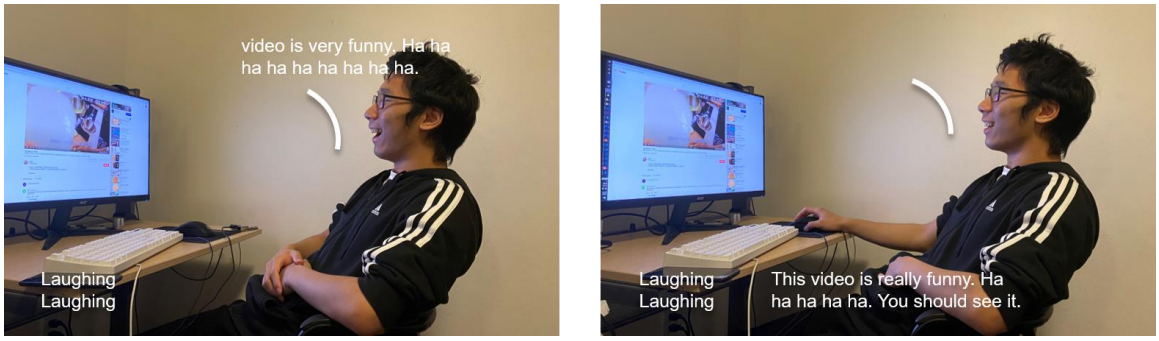


Figure 6.13: Speech-transcription can either be (A) placed on top of the speakers in the 3D space (*windows* view) or (B) move with the user's head (*subtitles* view).

6.4.3 Sound Recognition

Besides speech transcription, HoloSound shows the three most recently recognized sound events (*e.g.*, doorbell, knocking) at the bottom-left corner of the display (Figure 6.11). We use a deep learning-based sound classification engine running on a cloud that continually senses and processes audio in real-time.

To create the classification engine, we followed an approach similar to *HomeSound* [54], which uses transfer learning to adapt a deep CNN-based image classification model (VGG) for sound classification. This model achieved an overall accuracy of 84.9% on sounds recorded in the homes. To train the VGG model, we used sound clips of 19 common sound classes preferred by DHH people (*e.g.*, door knock, fire alarm, phone ring) from online sound effects libraries (*e.g.*, BBC [152], FreeSound [28]). All clips were converted to a single format (16KHz, 16-bit, mono) and silences greater than one second were removed, resulting in 27.8 hours of recordings. We then used the method in Hershey *et al.* [41] to compute the log-mel spectrogram features, which were fed to the model.

To process sounds in real-time, HoloSound uses a sliding window approach to sample the microphone at 16KHz (16,000 samples every second), extract the log-mel spectrogram features, and upload the 1-second buffer to the cloud. After classification, all sounds below 50% confidence and 45dB loudness are ignored.

6.4.4 Sound Localization

The third key desired property conveyed by HoloSound is sound location. For localization, we use ReSpeaker [171], an external portable 4-microphone array (Figure 6.12), which we ultimately envision could be integrated into future AR devices. Though HoloLens has four onboard microphones [173], these microphones

are specifically designed for voice input from the user (*e.g.*, by enhancing audio input from the user's face through beamforming) [3] and thus cannot be used for accurate localization. The ReSpeaker array is coupled to a Raspberry Pi 4 [174] running a modified 3D Kalman filter sound localization algorithm [36]. After processing, the direction of at most four sound sources in the user's vicinity is sent to the HoloSound app through a WIFI server. To visualize each sound, the continuous 3D direction is projected to one of 12 discrete directions in the horizontal plane; these values are then shown as circular arcs in a top-down view on the HoloLens display's vertical plane (Figure 6.11).

6.4.5 Study 4 Summary and Discussion

This section contributes the design and implementation of an augmented reality system, called *HoloSound*, that uses a head-mounted display and an external microphone array for transcribing speech, identifying sounds, and localizing sound sources, which are displayed to the user in the 3D space. However, considerable work remains in studying this system with DHH users and iterating on its design. Below, we discuss our future plans for a user study and further exploration opportunities for AR-based sound awareness.

UI exploration. Our system is open-sourced and can be used to prototype multiple UI designs and conduct a design probe study with DHH individuals. A specific goal of such a design probe could be to explore how the UI designs may vary with different social contexts (*i.e.*, 1:1 meeting *vs.* a midsize meeting *vs.* a large lecture). Specific research questions of interest, include:

1. In what contexts do the users desire full transcription vs. a topical summary?
2. How might the UI design vary with the conversation importance (*e.g.*, with friends vs. at workplace)?
3. How many source locations should be shown simultaneously, and how does this vary with context?
4. What non-speech sounds are desired for each context? How should these sounds be visualized?
5. Should the wearer's voice be transcribed?

A key aspect of this study will involve balancing cognitive load with useful information. For example, if the user is involved in an important 1:1 meeting, the system could prioritize speech transcription and show alerts for important non-speech sounds only (*e.g.*, fire alarms).

System exploration. Future work should also perform accuracy and performance (latency, CPU, and memory usage) testing of our sound recognition and localization systems, which could have a significant impact on the user experience. For localization accuracy, a sound source could be placed at different angles in front of our system while measuring the mean angular error and standard deviation. For sound recognition accuracy, since the accuracy may vary with audio contexts and background noise, a hearing user could wear the device in different physical locations (*e.g.*, home, in transit, office) for several hours and report on whether the sound recognition is accurate. Finally, for performance testing, future work could play recordings of real-life sounds and speech on a computer placed near our system and measure the HoloLens' CPU and memory usage as well as the end-to-end latencies of our speech transcription, sound recognition, and sound localization features.

Examining complementary haptic feedback. Another potential area of exploration is complementing the visual HMD feedback with haptic notifications delivered through a smartwatch or a custom hardware solution. While haptic feedback provides more limited bandwidth than visual feedback, it can be used to provide complementary information—such as to notify the user of an important non-speech sound [6], or to enhance transcription by providing speech tone [7]. Future work should compare performance tradeoffs in providing haptic notifications to complement visual information. One idea is to conduct a controlled study with varying device combinations (HMD-only, HMD + smartwatch, HMD + custom haptic hardware) and feedback modalities (visual-only, visual+haptic). Participants could be given a distractor task and asked to localize sounds emanating from a circular array of speakers placed around them, while measurements of speed and accuracy of identifying the sound source, as well as self-reported cognitive load are taken.

Chapter 7: Conclusion and Future Work

Taken together, the focus of this dissertation was to understand DHH people's needs and preferences for sound awareness, design technology interventions to support those needs, and evaluate those interventions with DHH users in naturalistic environments. Overall, my work has contributed novel sound sensing, processing, and visualization techniques, with an overarching vision to transform how DHH people think about, experience, and engage with the sound in their daily lives. In this chapter, I discuss the future implications of this dissertation. I shift to first-person language to highlight my own specific contributions.

7.1 DHH People and Technology Preferences

First off, an important note about DHH people and their technology preferences. My work is heavily informed by DHH perspectives and past work [54,55], but I do not assume it is universally desired or that it will necessarily work, as designed, for all users. As explained throughout this document, some DHH people may feel negatively towards some of the sound awareness technology, especially those who identify as part of the Deaf culture [15,71]. However, my surveys and evaluations provide strong evidence that many DHH individuals would find sound feedback valuable in their daily lives. I offer the technology artifacts not as mandatory interventions, but as "choices" for individual DHH people to use if they are unsatisfied with their current ways of sound awareness.

7.2 Contributions

Overall, my work advances the area of sound accessibility by:

1. Characterizing sound awareness needs and preferences of DHH users in multiple contexts
2. Designing four system artifacts: *HomeSound*, *SoundWatch*, *ProtoSound*, and *HoloSound* to enhance sound awareness in multiple contexts, and
3. Contributing insights from evaluations of these systems—which included the full spectrum from short, controlled evaluations to ecologically-valid field deployments.

More specifically,

For homes, I contributed insights from formative studies that examine the sound awareness needs of DHH people in the home, including the preferences for the sounds, visualizations, and reactions to specific themes that arise in a home context such as privacy, information overload, and effect on social dynamics. Based on these insights, I designed a smarthome system, called HomeSound, the first IoT-based sound awareness system for the home, and iteratively evaluated it using three-week field studies. The findings provide evidence of participants' change in understanding of sounds in their home over time, and their ability to perform essential daily tasks faster or with ease. I also uncovered several improvement suggestions including the need to accurately convey errors in sound recognition and mitigate information overall.

For portable environments, I contributed the design and evaluation of SoundWatch, the first smartwatch sound recognition system that uses deep learning to classify in diverse contexts. Through this work, I also contributed a performance comparison of several lightweight sound classification models and classification pipelines. User evaluations of SoundWatch suggest that it can help support sound awareness in multiple contexts; however, classification accuracy needed to be improved, especially for outdoor settings.

To further diversify the context of use, I contributed ProtoSound, a system to personalize a sound recognition model to support users' custom sounds and individual use cases. Through designing and evaluating ProtoSound, I contributed: several personalization preferences of DHH people gleaned from a survey with 472 participants, comparison of state-of-the-art machine learning personalization techniques on real-world sound datasets, and insights from the field evaluation of the most promising personalization technique.

Towards speech awareness, I contributed insights from: (1) a 45-day autoethnographic evaluation of initial head-mounted display (HMD)-based captioning interfaces, (2) a semi-controlled evaluation of refined HMD-based captioning interfaces, and (3) a full HMD-based system, called *HoloSound*, that displays three most preferred speech and sound cues—captioning, speaker location, and associated non-speech sounds—to support DHH users' conversations.

7.3 Future Technical Directions in Sound Accessibility

I envision a future where sounds are accessible to DHH people everywhere, enabling them to be more aware of their surroundings, easily access spoken conversation, and obtain critical information in diverse contexts. From a technological standpoint, achieving this vision will require advancements in sound sensing systems, sound feedback design, sound datasets, and AI technology.

Sound Processing Systems. Developing sound accessibility systems that work in diverse contexts and for multiple user groups is challenging due to varying acoustic conditions and user requirements. My ProtoSound work has shown that by subtly involving end-users in the model training pipeline, we can build personalized systems that adapt to user-specific requirements and contextual needs [34,56]. Such human-AI approaches combine the flexibility of humans with the computational power of AI to solve complex problems that are impossible with either alone. However, before these techniques can be widely adopted, many challenges remain such as *how to collect user data while preserving privacy? how to achieve the right balance between user agency and user effort? and what device configurations and combinations (e.g., smartwatch, smartphone, and/or stationary displays) are promising?* Future work should explore reinforcement learning techniques that can adapt the model to contextual needs *on-the-go* using minimal user effort. Such techniques may be able to capture even greater contextual and temporal variations of real-world acoustic events beyond possible, for example, through ProtoSound (*e.g.*, varying cries of a baby or different piano notes).

Sound Feedback Interfaces. Beyond processing sound information, unobtrusively conveying it to the end-users is key to achieving sound accessibility. My work has explored rich design spaces to convert sound to visual and haptic feedback [49,51,52]. However, many questions remain, which should be explored through controlled studies of future prototypes, such as *how to gain a user's trust by effectively conveying the confidence and error in sound processing? how to combine multiple sound information (e.g., sound type, location, duration) on a small emerging display (e.g., heads-up display, smartwatch)? how to convey the associated semantics or the intent of a sound (e.g., 'actionable' cues such as a microwave beep or a fire alarm vs. 'experiential' information such as a bird chirp or wind blowing)?*

Sound Datasets. AI offers a great promise to improve accessibility since it can automatically learn from new data, allowing adaptation to a variety of users across environments. However, AI relies on datasets that are rooted in human problems, are unbiased, and are collected from diverse contexts. Huge disparities exist in current sound datasets which are collected by specific populations in limited contexts. Future work should focus on increasing dataset diversity by engaging marginalized groups such as people with disabilities or the global south in the data collection process. This is a challenging task. Consider for example, *how would someone who is deaf collect high-quality sounds if they cannot hear them?* To address this, our team is researching cause-and-effect visualizations (*e.g.*, showing cluster visualizations of the classifier) in a parallel thread with an aim to make end-users understand how the quality of their recorded samples may shape the model [34]. A key concern with data collection is privacy—*e.g.,* deaf users may not know what sound data is collected about them—and interfaces need to be carefully designed to address this.

7.4 Examining Socio-Cultural Contexts of Technology Use

The primary strength of my work comes from the analysis of environmental, cultural, and social contexts surrounding my sound awareness technology use, as opposed to past evaluations where technology designs were evaluated out of contexts (*e.g.,* [80,89]) or through narrowly defined outcomes (*e.g.,* improved comprehension [105] or performance [112]). The grounded evaluations uncovered ecologically valid insights about the impact of the technology on people's lives (*e.g.,* their ability to perform household chores) which are otherwise hard to obtain through narrow, controlled evaluations. Future work should conduct longer evaluations in naturalistic settings, to continue to investigate concerns such as the technology's effect on users' privacy, social dynamics, and information overload in differing real-world contexts.

The DHH culture could also influence technology preferences. In general, DHH people belong to three distinct culture groups: (1) capital 'D' Deaf who rely on sign language and maintain specific cultural norms, (2) small 'd' deaf who primarily connect to deafness audiologically, and (3) hard-of-hearing who refrain from identifying with a distinct cultural group. As my work shows, these culture differences could influence technology adoption—for example, Deaf people had different notions of privacy than deaf or hard of hearing individuals. Future work should conduct controlled studies with a greater number of people from each group to verify our initial findings and uncover other differences in sound awareness preferences, if any.

Finally, some wearable technology systems I built (*e.g.*, SoundWatch) make the burden of access fall on DHH people, who need to carry and use these technologies. For its full effectiveness, accessibility should be a shared responsibility—hearing people should also be involved ensuring that access is successfully negotiated and provided [9,57,84]. Of course, technologies should continue to center DHH perspective—because if these technologies do not work, it is DHH people who lose access—but there may be ways to reframe sound awareness as a community-based accommodation. For example, a recent work by McDonnell *et al.* [84] seeks to position captioning as a group accommodation by developing interfaces that persuade hearing conversation peers to adopt DHH friendly speakers (*e.g.*, speak slowly, pause repeatedly). Based on the level of desired hearing intervention, various technology pipelines can be examined ranging from a simple interface to infrequently correct sound recognition errors to a fully crowdsourced sound recognition pipeline.

7.5 Closing Remark: Towards Complete Sound Awareness

My work in this dissertation attempted to provide general sound awareness to DHH people. The findings show that I was successful in enhancing awareness about some environmental sounds by showing sound recognition and other simple sound visualizations (*e.g.*, to convey loudness, pitch). However, complete sound awareness is more than that. Consider, for example, a kettle boiling scenario. A hearing person will hear the initial whistles from the kettle, anticipate the impending boiling, and will start moving towards the kettle. In contrast, a DHH person using the SoundWatch app—which only displays a "kettle boiling" sound one time towards the end—will not get the same semantic experience.

So, how do we move from conveying discrete sound cues to providing complete 'semantic' sound awareness? We need to start by asking DHH people what semantic meanings they want to derive from sounds. A Deaf person may not want to perceive the surrounding sonic environment in the same way a hearing person might (*e.g.*, they may prefer to use visual cues instead of sounds to understand that a fan may be running). Moreover, some profoundly deaf people may not have any perceptual basis for sound and inquiring about their meaning of sound awareness from them could be challenging. We should also ensure that we are not making such category of people—who may be satisfied with their non-auditory ways of living—feel inadequate by obtaining novel awareness about missing sounds.

References

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, and others. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.

[2] Jakob Abeßer. 2020. A review of deep learning based methods for acoustic scene classification. *Applied Sciences* 10, 6.

[3] Manish Sharma, Mallikarjuna Rao Abhijit Jana. *HoloLens Blueprints - Google Books*.

[4] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

[5] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. 2019. TAU Moving Sound Events 2019 - Ambisonic, Anechoic, Synthetic IR and Moving Source Dataset [Data set].

[6] Rosa Ma Alsina-Pagès, Joan Navarro, Francesc Al\'\ias, and Marcos Hervás. 2017. homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. *Sensors* 17, 4: 854.

[7] Edward T. Auer. 1998. Temporal and spatio-temporal vibrotactile displays for voice fundamental frequency: An initial evaluation of a new vibrotactile speech perception aid with normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America* 104, 4: 2477.

[8] Sanghamitra Bandyopadhyay and Sriparna Saha. 2012. *Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications*. Springer Science & Business Media.

[9] Cynthia L Bennett, Erin Brady, and Stacy M Branham. 2018. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 161–173.

[10] Hugh Beyer and Karen Holtzblatt. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

[11] John W Du Bois and Stephan Schuetze-Coburn. 1993. Representing hierarchy: Constituent structure for discourse databases. *Talking data: Transcription and coding in discourse research*: 221–259.

[12] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12: 2301–2309.

[13] Danielle Bragg, Nicholas Huynh, and Richard E. Ladner. 2016. 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*, 3–13.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2: 77–101.

[15] Anna Cavender and Richard E Ladner. 2008. Hearing impairments. In *Web accessibility*. Springer, 25–35.

[16] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

[17] Jesper Dammeyer, Christine Lehane, and Marc Marschark. 2017. Use of technological aids and interpretation services among children and adults with hearing loss. *International journal of audiology* 56, 10: 740–748.

[18] Mohammad I Daoud, Mahmoud Al-Ashi, Fares Abawi, and Ala Khalifeh. 2015. In-house alert

sounds detection and direction of arrival estimation to assist people with hearing difficulties. In *Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on*, 297–302.

[19] Marilyn E. Demorest and Sue Ann Erdman. 1986. Scale Composition and Item Analysis of the Communication Profile for the Hearing Impaired. *Journal of Speech Language and Hearing Research* 29, 4: 515–535.

[20] Marilyn E. Demorest and Sue Ann Erdman. 1987. Development of the Communication Profile for the Hearing Impaired. *Journal of Speech and Hearing Disorders* 52, 2: 129–143.

[21] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2: 1–37.

[22] Carolyn Ellis, Tony E Adams, and Arthur P Bochner. 2011. Autoethnography: an overview. *Historical Social Research/Historische Sozialforschung*: 273–290.

[23] Johan Engström, Nina Åberg, Emma Johansson, and Jakob Hammarbäck. 2005. Comparison between visual and tactile signal detection tasks applied to the safety assessment of in-vehicle information systems. In *Driving Assesment Conference*.

[24] Michael Erard. 2017. Why Sign-Language Gloves Don't Help Deaf People. *The Atlantic*.

[25] Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and Hard-of-hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies. In *SIGCHI Conference on Human Factors in Computing Systems (CHI).*, 1–13.

[26] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135.

[27] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* 65: 22–28.

[28] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.*

[29] Karyn L. Galvin, Jan Ginis, Robert S. Cowan, Peter J. Blamey, and Graeme M. Clark. 2001. A Comparison of a New Prototype Tickle Talker™ with the Tactaid 7. *Australian and New Zealand Journal of Audiology* 23, 1: 18–36.

[30] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

[31] Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen. 2018. Unsupervised adversarial domain adaptation for acoustic scene classification. *arXiv preprint arXiv:1808.05777*.

[32] Abraham Glasser, Kesavan Kushalnagar, and Raja Kushalnagar. 2017. Deaf, hard of hearing, and hearing perspectives on using automatic speech recognition in conversation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 427–432.

[33] Steven Goodman, Susanne Kirchner, Rose Guttman, Dhruv Jain, Jon Froehlich, and Leah Findlater. 2020. 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*, 1–

13.

[34] Steven M Goodman, Ping Liu, Dhruv Jain, Emma J McDonnell, Jon E Froehlich, and Leah Findlater. 2021. Toward User-Driven Sound Recognizer Personalization with People Who Are d/Deaf or Hard of Hearing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2: 1–23.

[35] Benjamin M Gorman. 2014. VisAural: a wearable sound-localisation device for people with impaired hearing. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, 337–338.

[36] François Grondin and François Michaud. 2019. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robotics and Autonomous Systems* 113: 63–80.

[37] Ru Guo, Yiru Yang, Johnson Kuang, Xue Bin, Dhruv Jain, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. HoloSound: Combining Speech and Sound Identification for Deaf or Hard of Hearing Users on a Head-mounted Display. In *ACM SIGACCESS Conference on Computers and Accessibility*, 1–4.

[38] Richard S Hallam and Roslyn Corney. 2014. Conversation tactics in persons with normal hearing and hearing-impairment. *International journal of audiology* 53, 3: 174–81.

[39] Seungyeop Han, Haichen Shen, Matthai Philipose, Sharad Agarwal, Alec Wolman, and Arvind Krishnamurthy. 2016. Mcdnn: An approximation-based execution framework for deep stream processing under resource constraints. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, 123–136.

[40] Rebecca Perkins Harrington and Gregg C Vanderheiden. 2013. Crowd caption correction (ccc). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, 45.

[41] Shawn Hershey, Sourish Chaudhuri, Daniel P W Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 131–135.

[42] Yasamin Heshmat, Carman Neustaedter, and Brendan DeBrincat. 2017. The Autobiographical Design and Long Term Usage of an Always-On Video Recording System for the Home. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (DIS '17), 675–687.

[43] Eric G Hintz, Michael D Jones, M Jeannette Lawler, Nathan Bench, and Fred Mangrubang. 2015. Adoption of ASL classifiers as delivered by head-mounted displays in a planetarium show. *Journal of Astronomy & Earth Sciences Education (JAESE)* 2, 1: 1–16.

[44] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[45] Shahidul Islam, William G Buttlar, Roberto G Aldunate, and William R Vavrik. 2014. Measurement of pavement roughness using android-based smartphone application. *Transportation Research Record* 2457, 1: 30–38.

[46] Dhruv Jain, Bonnie Chinh, Leah Findlater, Raja Kushalnagar, and Jon Froehlich. 2018. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. In *ACM Conference Companion Publication on Designing Interactive Systems*, 7–11.

[47] Dhruv Jain, Brendon Chiu, Steven Goodman, Chris Schmandt, Leah Findlater, and Jon E Froehlich. 2020. Field study of a tactile sound awareness device for deaf users. In *Proceedings of the 2020 International Symposium on Wearable Computers*, 55–57.

[48] Dhruv Jain, Audrey Desjardins, Leah Findlater, and Jon E Froehlich. 2019. Autoethnography of a Hard of Hearing Traveler. In *The 21st International ACM SIGACCESS Conference on Computers*

and Accessibility, 236–248.

[49] Dhruv Jain, Leah Findlater, Christian Volger, Dmitry Zotkin, Ramani Duraiswami, and Jon Froehlich. 2015. Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing. In *ACM Conference on Human Factors in Computing Systems*, 241–250.

[50] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People Who are Deaf and Hard of Hearing. In *ASSETS 2018 - Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 81–92.

[51] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John R Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. Towards Sound Accessibility in Virtual Reality. In *ACM International Conference on Multimodal Interaction*, 1–15.

[52] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. A Taxonomy of Sounds in Virtual Reality. In *Designing Interactive Systems Conference 2021*, 160–170.

[53] Dhruv Jain, Angela Carey Lin, Marcus Amalachandran, Aileen Zeng, Rose Guttman, Leah Findlater, and Jon Froehlich. 2019. 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*, 94:1-94:13.

[54] Dhruv Jain, Kelly Mack, Akli Amrous, Matt Wright, Steven Goodman, Leah Findlater, and Jon E Froehlich. 2020. 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), 1–12.

[55] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. 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*, 1–13.

[56] Dhruv Jain, Khoa Nguyen, Steven Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon Froehlich. 2021. ProtoSound: A Personalized, Scalable Sound Recognition System for d/Deaf and Hard of Hearing Users. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 1–16.

[57] Dhruv Jain, Venkatesh Potluri, and Ather Sharif. 2020. Navigating Graduate School with a Disability. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–11.

[58] Charlene A Johnson. 2010. Articulation of Deaf and Hearing Spaces Using Deaf Space Design Guidelines: A Community Based Participatory Research with the Albuquerque Sign Language Academy.

[59] Michael Jones, M Jeannette Lawler, Eric Hintz, Nathan Bench, Fred Mangrubang, and Mallory Trullender. 2014. Head Mounted Displays and Deaf Children: Facilitating Sign Language in Challenging Learning Environments. In *Proceedings of the 2014 Conference on Interaction Design and Children* (IDC '14), 317–320.

[60] Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio A B Penatti, Ricardo da S Torres, and Anderson Rocha. 2017. Nearest neighbors distance ratio open-set classifier. *Machine Learning* 106, 3: 359–386.

[61] Y Kaneko, Inho Chung, and K Suzuki. 2013. Light-Emitting Device for Supporting Auditory Awareness of Hearing-Impaired People during Group Conversations. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 3567–3572.

[62] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM*

SIGARCH Computer Architecture News 45, 1: 615–629.

[63] Mahdie Karbasi, Seyed Mohammad Ahadi, and M Bahmanian. 2011. Environmental sound classification using spectral dynamic features. In *2011 8th International Conference on Information, Communications & Signal Processing*, 1–5.

[64] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[65] W Bradley Knox and Peter Stone. 2015. Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. *Artificial Intelligence* 225: 24–50.

[66] Sergei Kochkin. 2000. MarkeTrak V:"Why my hearing aids are in the drawer" The consumers' perspective. *The Hearing Journal* 53, 2: 34–36.

[67] Ines Kožuh, Manfred Hintermair, and Matjaž Debevc. 2016. Community building among deaf and hard of hearing people by using written language on social networking sites. *Computers in Human Behavior* 65: 295–307.

[68] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[69] Raja. S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Transactions on Accessible Computing* 5, 3: 1–24.

[70] Jim G Kyle and Bencie Woll. 1988. *Sign language: The study of deaf people and their language*. Cambridge University Press.

[71] Paddy Ladd and Harlan Lane. 2013. Deaf ethnicity, deafhood, and their relationship. *Sign Language Studies* 13, 4: 565–579.

[72] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play

acoustic activity recognition. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, 213–224.

[73] Walter S Lasecki, Christopher D Miller, Raja S Kushalnagar, and Jeffrey P Bigham. 2013. Legion Scribe: Real-Time Captioning by the Non-Experts. In *10th International Cross-Disclipinary Conference on Web Accessibility (W4A)*.

[74] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, and Samarjit Das. 2017. A comparison of deep learning methods for environmental sound detection. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 126–130.

[75] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. 2002. Content analysis for audio classification and segmentation. *IEEE Transactions on speech and audio processing* 10, 7: 504–516.

[76] Michella Maiorana-Basas and Claudia M Pagliaro. 2014. Technology use among adults who are deaf and hard of hearing: A national survey. *Journal of deaf studies and deaf education* 19, 3: 400–410.

[77] Manuela M Marin and Helmut Leder. 2013. Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PloS one* 8, 8: e72412.

[78] Tara Matthews. 2006. Designing and Evaluating Glanceable Peripheral Displays. In *Proceedings of the 6th Conference on Designing Interactive Systems* (DIS '06), 343–345.

[79] Tara Matthews, Scott Carter, Carol Pai, Janette Fong, and Jennifer Mankoff. 2006. Scribe4Me: Evaluating a Mobile Sound Transcription Tool for the Deaf. In *Proceedings of Ubiquitous Computing (UbiComp)*, Paul Dourish and Adrian Friday (eds.). Springer Berlin Heidelberg, 159–176.

[80] Tara Matthews, Janette Fong, F. Wai-Ling Ho-Ching, and Jennifer Mankoff. 2006. Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology* 25, 4: 333–351.

[81] Tara Matthews, Janette Fong, and Jennifer Mankoff. 2005. Visualizing non-speech sounds for the deaf. In *ACM SIGACCESS conference on Computers and accessibility*, 52.

[82] Amrita Mazumdar, Brandon Haynes, Magda Balazinska, Luis Ceze, Alvin Cheung, and Mark Oskin. 2019. Perceptual Compression for Video Storage and Processing Systems. In *Proceedings of the ACM Symposium on Cloud Computing*, 179–192.

[83] Abby McCormack and Heather Fortnum. 2013. Why do people fitted with hearing aids not wear them? *International Journal of Audiology* 52, 5: 360–368.

[84] Emma J McDonnell, Ping Liu, Steven M Goodman, Raja Kushalnagar, Jon E Froehlich, and Leah Findlater. 2021. Social, environmental, and technical: Factors at play in the current use and future design of small-group captioning. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2: 1–25.

[85] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT Sound events 2016. https://doi.org/10.5281/zenodo.45759

[86] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2019. Acoustic Scene Classification in DCASE 2019 Challenge: Closed and Open Set Classification and Data Mismatch Setups. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 164–168.

[87] Michael Potuck. Hands-on with iOS 14's Sound Recognition feature that listens for doorbells, smoke alarms, more. Retrieved March 9, 2021 from https://9to5mac.com/2020/10/28/how-to-use-iphone-sound-recognition-ios-14/

[88] Matthias Mielke and Rainer Brück. 2015. A Pilot Study about the Smartwatch as Assistive Device for Deaf People. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, 301–302.

[89] Matthias Mielke and Rainer Brueck. 2015. Design and evaluation of a smartphone application for non-speech sound awareness for people with hearing loss. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 5008–5011.

[90] Ashley Miller, Joan Malasig, Brenda Castro, Vicki L Hanson, Hugo Nicolau, and Alessandra Brandão. 2017. The Use of Smart Glasses for Lecture Comprehension by Deaf and Hard of Hearing Students. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17), 1909–1915.

[91] Matthew S Moore. 1992. *For Hearing people only: Answers to some of the most commonly asked questions about the Deaf community, its culture, and the" Deaf Reality"*. Deaf Life Press.

[92] National Association of the Deaf (NAD). Communication Access Realtime Translation. Retrieved April 7, 2018 from https://www.nad.org/resources/technology/captioning-for-access/communication-access-realtime-translation/

[93] National Institute of Deafness and Other Communication Disorders. 2016. Quick Statistics About Hearing. Retrieved September 3, 2018 from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing

[94] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

[95] Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, 86–93.

[96] Boris N Oreshkin, Pau Rodr\'\iguez López, and Alexandre Lacoste. 2018. {TADAM:} Task dependent adaptive metric for improved few-shot learning. *CoRR* abs/1805.1. Retrieved from http://arxiv.org/abs/1805.10123

[97] Phil Parette and Marcia Scherer. 2004. Assistive Technology Use and Stigma. *Education and Training in Developmental Disabilities-September 2004*: 217–226.

[98] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

[99] Yi-Hao Peng, Ming-Wei Hsu, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, Paper No. 293.

[100] Benjamin Petry, Thavishi Illandara, Don Samitha Elvitigala, and Suranga Nanayakkara. 2018. Supporting Rhythm Activities of Deaf Children Using Music-Sensory-Substitution Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 486:1--486:10.

[101] Silvia Pfeiffer, Stephan Fischer, and Wolfgang Effelsberg. 1997. Automatic audio content analysis. In *Proceedings of the fourth ACM international conference on Multimedia*, 21–30.

[102] A J Phillips, A R D Thornton, S Worsfold, A Downie, and J Milligan. 1994. Experience of using vibrotactile aids with the profoundly deafened. *European journal of disorders of communication* 29, 1: 17–26.

[103] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1015–1018.

[104] Mary R Power, Des Power, and Louise Horstmanshof. 2006. Deaf people communicating via SMS, TTY, relay service, fax, and computers in Australia. *Journal of deaf studies and deaf education* 12, 1: 80–92.

[105] Adele Proctor. 1990. Oral language comprehension using hearing aids and tactile aids: Three case studies. *Language, Speech, and Hearing Services in Schools* 21, 1: 37–48.

[106] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2016)*, 10 pages.

[107] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. 2018. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7229–7238.

[108] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. 2019. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*.

[109] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human--Computer Interaction* 35, 5–6: 413–451.

[110] Dhrubojyoti Roy, Sangeeta Srivastava, Aditya Kusupati, Pranshu Jain, Manik Varma, and Anish Arora. 2019. One size does not fit all: Multi-scale, cascaded RNNs for radar classification. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 1–10.

[111] J Salamon, C Jacoby, and J P Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *22nd {ACM} International Conference on Multimedia (ACM-MM'14)*, 1041–1044.

[112] Frank A Saunders, William A Hill, and Barbara Franklin. 1981. A wearable tactile sensory aid for profoundly deaf children. *Journal of Medical Systems* 5, 4: 265–270.

[113] John Saunders. 1996. Real-time discrimination of broadcast speech/music. In *1996 IEEE*

International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 993–996.

[114] Eric Scheirer and Malcolm Slaney. 1997. Construction and evaluation of a robust multifeature speech/music discriminator. In *1997 IEEE international conference on acoustics, speech, and signal processing*, 1331–1334.

[115] Khurram Shahzad and Bengt Oelmann. 2014. 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)*, 519–524.

[116] Bowen Shi, Ming Sun, Krishna C Puvvada, Chieh-Chi Kao, Spyros Matsoukas, and Chao Wang. 2020. Few-shot acoustic event detection via meta learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 76–80.

[117] Kristen Shinohara and Jacob O Wobbrock. 2016. Self-Conscious or Self-Confident? A Diary Study Conceptualizing the Social Accessibility of Assistive Technology. *ACM Trans. Access. Comput.* 8, 2: 5:1--5:31.

[118] Kristen Shinohara and JO Wobbrock. 2011. In the shadow of misperception: assistive technology use and social interactions. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 705–714.

[119] Liu Sicong, Zhou Zimu, Du Junzhao, Shangguan Longfei, Jun Han, and Xin Wang. 2017. UbiEar: Bringing Location-independent Sound Awareness to the Hard-of-hearing People with Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2: 17.

[120] Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

[121] Ole Morten Strand and Andreas Egeberg. 2004. Cepstral mean and variance normalization in the

model domain. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*.

[122] Kazuki Suemitsu, Keiichi Zempo, Koichi Mizutani, and Naoto Wakatsuki. 2015. Caption support system for complementary dialogical information using see-through head mounted display. In *Consumer Electronics (GCCE), 2015 IEEE 4th Global Conference on*, 368–371.

[123] I R Summers, M A Peake, and M C Martin. 1981. Field trials of a tactile acoustic monitor for the profoundly deaf. *British journal of audiology* 15, 3: 195–199.

[124] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 403–412.

[125] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[126] Sasha Targ, Diogo Almeida, and Kevin Lyman. 2016. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

[127] Stefan Tilkov and Steve Vinoski. 2010. Node. js: Using JavaScript to build high-performance network programs. *IEEE Internet Computing* 14, 6: 80–83.

[128] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 242–264.

[129] Mike Wald. 2006. Captioning for Deaf and Hard of Hearing People by Editing Automatic Speech Recognition in Real Time. In *Proceedings of the 10th International Conference on Computers Helping People with Special Needs* (ICCHP'06), 683–690.

[130] Mike Wald and Keith Bain. 2007. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society* 6, 4: 435–447.

[131] Wei Wang and Zhi-Hua Zhou. 2010. A new analysis of co-training. In *ICML*.

[132] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* 53, 3: 1–34.

[133] Yu Wang, Justin Salamon, Nicholas J Bryan, and Juan Pablo Bello. 2020. Few-shot sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 81–85.

[134] Samuel White. 2010. Audiowiz: Nearly Real-time Audio Transcriptions. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '10), 307–308.

[135] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, and others. 2019. Disability, bias, and AI. *AI Now Institute*.

[136] Jason Wu, Chris Harrison, Jeffrey P Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

[137] Nobuhide Yamakawa, Tetsuro Kitahara, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. 2010. Effects of modelling within-and between-frame temporal variations in power spectra on non-verbal sound recognition. In *Eleventh Annual Conference of the International Speech Communication Association*.

[138] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2:

1–19.

[139] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*.

[140] Eddy Yeung, Arthur Boothroyd, and Cecil Redmond. 1988. A Wearable Multichannel Tactile Display of Voice Fundamental Frequency. *Ear and Hearing* 9, 6: 342–350.

[141] Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*.

[142] Hanfeng Yuan, Charlotte M. Reed, and Nathaniel I. Durlach. 2005. Tactual display of consonant voicing as a supplement to lipreading. *The Journal of the Acoustical Society of America* 118, 2: 1003.

[143] Alina Zajadacz. 2015. Evolution of models of disability as a basis for further policy changes in accessible tourism. *Journal of Tourism Futures* 1, 3: 189–202.

[144] Quick Statistics About Hearing | NIDCD. Retrieved July 7, 2022 from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing

[145] Live Transcribe | Speech to Text App | Android. Retrieved August 12, 2022 from https://www.android.com/accessibility/live-transcribe/

[146] What is real-time captioning? | UW DO-IT. Retrieved August 12, 2022 from https://www.washington.edu/doit/what-real-time-captioning#:~:text=Captions%2C composed of text%2C are,as an event takes place.

[147] DeafSpace – Gallaudet University. Retrieved August 12, 2022 from https://www.gallaudet.edu/campus-design-and-planning/deafspace/

[148] EventDrops: A time based / event series interactive visualization using d3.js. Retrieved August 12, 2022 from https://www.npmjs.com/package/eventDrops

[149] Luke Knox - Visualoop. Retrieved September 15, 2019 from http://visualoop.com/blog/33248/portfolio-of-the-week-luke-knox

[150] PyAudio. Retrieved September 15, 2019 from https://people.csail.mit.edu/hubert/pyaudio/

[151] PM2 - Advanced Node.js process manager. Retrieved September 15, 2019 from http://pm2.keymetrics.io/

[152] BBC Sound Effects. Retrieved September 18, 2019 from http://bbcsfx.acropolis.org.uk/

[153] Network Sound Effects Library. Retrieved September 15, 2019 from https://www.sound-ideas.com/Product/199/Network-Sound-Effects-Library

[154] UPC-TALP dataset. Retrieved September 18, 2019 from http://www.talp.upc.edu/content/upc-talp-database-isolated-meeting-room-acoustic-events

[155] Mobvoi TicWatch E2. Retrieved September 15, 2019 from https://www.mobvoi.com/us/pages/ticwatche2

[156] Hosted models | TensorFlow Lite. Retrieved May 5, 2020 from https://www.tensorflow.org/lite/guide/hosted_models

[157] TicWatch Pro - Mobvoi. Retrieved May 5, 2020 from https://www.mobvoi.com/au/pages/ticwatchpro

[158] Honor 7X - Huawei. Retrieved May 5, 2020 from https://www.gsmarena.com/honor_7x-8880.php

[159] AI Inference: Applying Deep Neural Network Training. Retrieved May 5, 2020 from https://mitxpc.com/pages/ai-inference-applying-deep-neural-network-training

[160] Measure app performance with Android Profiler | Android Developers. Retrieved May 5, 2020 from https://developer.android.com/studio/profile/android-profiler

[161] Profile battery usage with Batterystats and Battery Historian. Retrieved May 5, 2020 from https://developer.android.com/topic/performance/power/setup-battery-historian

[162] General Data Protection Regulation (GDPR) – Official Legal Text. Retrieved July 21, 2020 from https://gdpr-info.eu/

[163] California Consumer Privacy Act (CCPA) | State of California - Department of Justice - Office of the Attorney General. Retrieved July 21, 2020 from https://gdpr-info.eu/

[164] Live Transcribe & Sound Notifications – Apps on Google Play. Retrieved April 5, 2021 from https://play.google.com/store/apps/details?id=com.google.audio.hearing.visualization.accessibility.scribe

[165] Google Surveys. Retrieved April 5, 2021 from https://surveys.google.com

[166] Google Opinion Rewards - It Pays to Share Your Opinion. Retrieved April 5, 2021 from https://surveys.google.com/google-opinion-rewards/

[167] Google AI Blog: Federated Learning: Collaborative Machine Learning without Centralized Training Data. Retrieved April 6, 2021 from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[168] AudioSet Label Accuracy. Retrieved April 6, 2021 from https://research.google.com/audioset/dataset/index.html

[169] TextHear Speech To Text Technologies for the Hearing Impaired. Retrieved July 4, 2022 from https://texthear.com/

[170] StreamText.Net. Retrieved March 7, 2018 from http://www.streamtext.net/

[171] ReSpeaker Mic Array v2.0 - Seed Wiki. Retrieved May 3, 2021 from https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

[172] Speech to Text | Microsoft Azure. Retrieved August 10, 2022 from https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-to-text

[173] HoloLens (1st gen) hardware | Microsoft Docs. Retrieved August 1, 2022 from https://docs.microsoft.com/en-us/hololens/hololens1-hardware

[174] Raspberry Pi 4. Retrieved August 2, 2022 from https://www.raspberrypi.com/products/raspberry-pi-4-model-b/