Human-Centered Sound Recognition Tools for Deaf and Hard of Hearing People

Steven Goodman

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Leah Findlater, Chair

Jon E. Froehlich
Julie Kientz

Program Authorized to Offer Degree:
Human Centered Science and Engineering

University of Washington

Abstract

Human-Centered Sound Recognition Tools for Deaf and Hard of Hearing People

Steven Goodman

Chair of the Supervisory Committee:

Associate Professor Leah Findlater
Human Centered Design and Engineering

Sound carries rich information about our surroundings; however, this information can be inaccessible to people who are Deaf, deaf, or hard of hearing (DHH). Automatic sound recognition features are now available on smartphones and other devices, but current implementations offer limited personalization options, hindering their ability to accommodate DHH users' diverse interests and varied contexts of use. In this dissertation, I present a series of iterative studies that explore the design of human-centered sound recognition tools to enable DHH users to tailor sound information to their individual needs. I examine user experiences across different stages of a machine learning workflow for personalization, including: problem framing for contextual sound feedback, capturing and curating personal audio data, and interactive training and evaluation of custom sound recognition models. Together, this work provides a comprehensive, empirical investigation into the challenges and opportunities with developing human-centered sound recognition tools for DHH users. I close with recommendations for interface design and opportunities for future research in this space.

Acknowledgements

Many individuals had a hand in the completion of this work, and I owe a debt of gratitude to each of them. First and foremost, I must thank my advisor, Leah Findlater: her instruction, feedback, patience, and insightful counsel were instrumental in carrying out this research. I appreciate her placing her trust in me as her first HCDE graduate student, and her mentorship has been invaluable in my training to become a thoughtful and articulate researcher. I will carry the lessons I've learned under her guidance throughout the remainder of my career.

Next, my collaborators: I extend sincere gratitude to Jon Froehlich, my enduring collaborator and eventual committee member, for his direction, assistance, and infectious enthusiasm toward this work. I am also indebted to my other co-authors, Dhruv Jain, Emma McDonnell, Susanne Kirchner, Rose Guttman, and Ping Liu, who each made invaluable contributions to this research. Thank you also to Raja Kushalnagar, Augustina Liu, Lucy Jiang, and Avery Mack for your support.

To the other committee members: thank you to Julie Kientz for your helpful advice, recommendations from throughout the HCI landscape, and keen interest in my work. To Mark Harniss, my GSR, thank you for providing suggestions for relevant Deaf Studies literature and your continued engagement with this research. Each member of my outstanding committee helped to shape and refine this dissertation, and I am lucky to have had your input.

I must acknowledge the members of the Inclusive Design Lab, both past and present, for creating such an enriching research environment. I especially want to thank Lotus Zhang, Gina Clepper, and Abigale Stangl for their fellowship, insightful discussions, and willingness to assist whenever it was needed. Though

from other labs, I must also mention Calvin Liang and Hannah Twigg-Smith, who provided much-needed camaraderie while I found my bearings early on.

I am grateful to the faculty and staff of HCDE for fostering the welcoming community that I was proud to call my academic home for the last six years. I especially want to thank Kathleen Rascon, Pat Reilly, Stacia Green, Allen Lee, and Leah Pistorius, along with Brock Craft, Sean Munson, and Jennifer Turns. Their assistance and generosity with their time were invaluable for navigating all of the unexpected quirks of a doctorate degree.

To my siblings, Meghan, Ben, Nolan, Jonathan, and Colin: thank you so much for your love, encouragement, and backing during my pursuit of this degree. While it kept me from seeing you and your families in person more than I would have liked, your efforts to connect with me always brought me levity and joy during the challenging times; I look forward to making up for that lost time going forward.

And lastly...

DEDICATION

To Katherine, my wonderful partner,
without whose unwavering belief, patience,
and support, this work would not exist.

And to my parents, Gary and Mary Pat—
you instilled in me the diligence and resilience
required to bring this work to fruition.

Contents

1 **Introduction** **11**

1.1 Thesis Statement 13

1.2 Document Outline 14

1.3 Summary of Contributions 16

1.4 Authorship Statement 17

2 **Background & Related Work** **19**

2.1 Disability Models and DHH Perspectives 19

2.2 Sound Awareness Needs and Preferences 21

2.3 Sound Awareness Formats and Feedback 22

2.4 Sound Recognition Tools 23

2.5 Human-Centered ML for Accessibility 25

3 **Evaluating Sound Feedback Across Contexts** **27**

3.1 Introduction 27

3.2 Method 29

3.3 Findings 36

3.4 Discussion 46

3.5 Chapter Summary 49

4 **Investigating Real-World Audio Data Collection** **51**

4.1 Introduction 51

4.2 Methods 54

4.3 Findings 61

4.4 Discussion 74

4.5 Chapter Summary 79

5 Evaluating an Accessible Sound Recognition System 81

5.1 Introduction 81

5.2 SPECTRA: A DHH-Centered Pipeline for Personalized Sound Models 83

5.3 Evaluation 92

5.4 Findings 97

5.5 Discussion 106

5.6 Chapter Summary 110

6 Conclusion & Future Work 111

6.1 Empirical Contributions 111

6.2 Design Guidelines for Future Personalizable Sound Recognition Tools 113

6.3 Opportunities for Future Work 115

6.4 Limitations 117

6.5 Closing Remark: Towards Transparency in Accessible AI 119

Chapter 1

Introduction

Sound carries rich information about the world around us, ranging from subtle cues like the rustling of leaves, household activities like a boiling tea kettle, or critical alerts like an approaching vehicle. However, this information can go unnoticed or be inaccessible to individuals who are Deaf, deaf, or hard of hearing (DHH). Prior work shows that many DHH people desire sound information to support personal safety (footsteps), social awareness (nearby voices), and attending to non-urgent alerts (home appliances) [13, 37, 70, 98]. A survey of 201 DHH participants found that smartwatches are the most preferred portable format for sound awareness due to their capability to provide glanceable, socially acceptable, and multimodal (visual and haptic) sound feedback [37]. Most prior work, however, has sought to build systems that provide sound information through a single modality (*e.g.*, [52, 64, 76]), and the effective delivery of sound feedback within this format—especially in complex and varied soundscapes—is an open design question.

The value of sound information is also strongly influenced by an individual's circumstances. For example, hard of hearing users have more interest in certain sounds (phone ringing, spoken conversations) than users identifying as Deaf or deaf [13, 37]. The relevance of sound information may also change as the user moves between social contexts (family vs. strangers) [13, 37, 65] and physical locations (at home vs. while mobile) [98]. Further, DHH users are commonly interested in distinctive sounds from their personal lives (children [98], name calls [13]) or their homes [71]. With preferences shaped by various cultural, social, locational, and personal factors, a "one-size-fits-all" sound awareness solution may not exist—rather, personalization options are needed to cater to the diverse preferences of DHH users.

Following advances in signal processing and machine learning (ML), sound recognition tools have proliferated, both in the research literature (*e.g.*, [13, 71, 127]) and in commercial applications—Android and iOS smartphones now include accessibility features to recognize common sounds such as doorbells, running water, and dog barks. However, these sound recognition features are built on generic models that offer limited flexibility for accommodating individual needs and diverse sound environments. Indeed, a recent survey of Android sound recognition users highlighted their dissatisfaction with the feature's limited sound categories and accuracy [67] while echoing prior work showing DHH users are interested in personalizing sound recognition systems themselves [13, 60, 72, 107]. Allowing users to engage with the ML pipeline can provide transparency and a sense of control [29], which can positively impact trust, satisfaction, and continued use of these systems [3, 83]. However, another open question lies in how to effectively support a DHH user—who does not have full access to a sound themselves—in capturing and selecting suitable audio data to train a machine learning (ML) model.

To address these challenges and explore the design space of *human-centered sound recognition tools*, my dissertation sought to answer the following core research questions:

RQ1. How do DHH users desire sound information to be delivered, and how do contextual factors impact these preferences?

RQ2. How do DHH users capture, interpret, and conceptualize audio data for automatic sound recognition?

RQ3. What feedback mechanisms and UI elements can effectively support DHH users to personalize a sound recognition system?

RQ4. How does a complete training and testing cycle affect DHH users' understanding and assessment of a personalizable sound recognition system?

1.1 Thesis Statement

My thesis statement is:

> *For DHH people who desire greater access to sound information, technology should be designed for personalized and adaptable experiences—providing relevant information, offering granular control, and promoting confidence and agency among users.*

To explore this thesis, I took a comprehensive, empirical approach to examine the needs and preferences of DHH users across different stages of an ML workflow for sound recognition, including: problem framing for sound feedback priorities and contextual use; capturing and curating personal audio datasets; and the interactive training and evaluation of a personalized model (Figure 1.1).

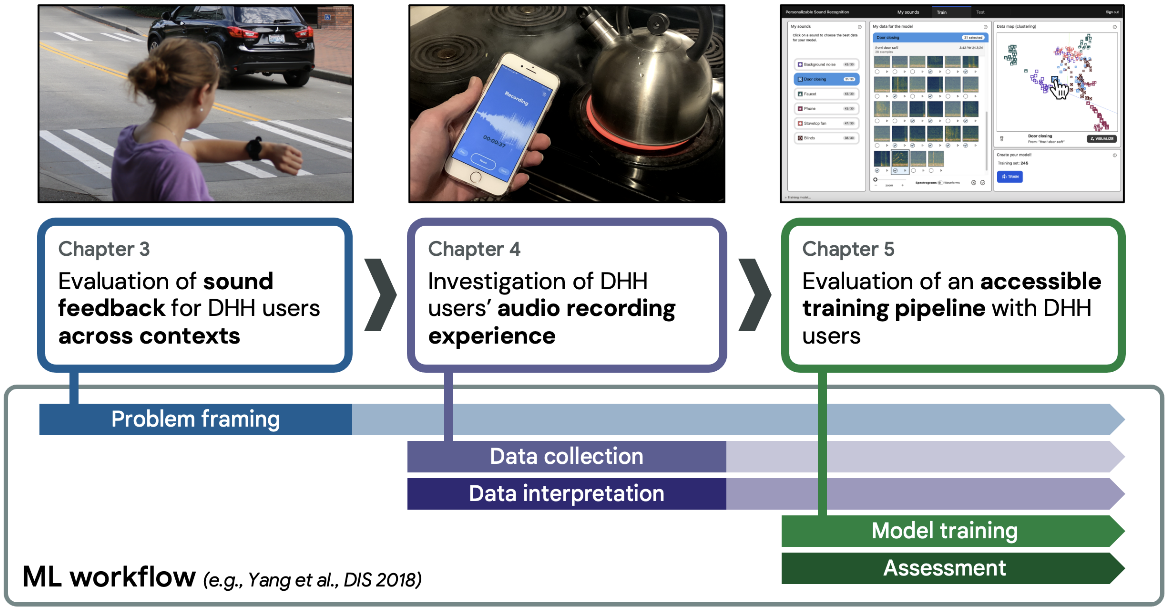

Chapter 3
Evaluation of **sound feedback** for DHH users **across contexts**
Chapter 4
Investigation of DHH users' **audio recording experience**
Chapter 5
Evaluation of an **accessible training pipeline** with DHH users
Problem framing
Data collection
Data interpretation
Model training
Assessment
ML workflow (e.g., Yang et al., DIS 2018)

Figure 1.1: My research explores the design space of human-centered sound recognition tools for DHH users across an ML workflow. It includes (Ch. 3) a smartwatch-based design probe into their contextual needs and feedback preferences, (Ch. 4) a field study of personal audio data collection, and (Ch. 5) the design and evaluation of a prototype to support interactive model training.

1.2 Document Outline

To address the core research questions outlined above, this dissertation is divided into the following chapters. Each chapter delves into a specific aspect of the user experience with personalizable sound recognition technology, yielding insights into DHH users' needs and preferences, guidelines for future personalizable tools, and key opportunities for future work.

Exploring Sound Feedback Across Contexts. The first study, detailed in Chapter 3, explored DHH users' preferences for receiving sound information and the impact of contextual factors on these preferences (RQ1). A large-scale survey of DHH participants highlighted smartwatches as the preferred portable device for sound awareness [37], but prior work has not explored effective ways of conveying sound information on these devices. This study aimed to fill this gap by examining the combination of visual and haptic feedback on smartwatches and exploring how sound filtering should be designed to accommodate different soundscapes.

I recruited 16 DHH participants for a three-part Wizard-of-Oz study with a smartwatch prototype [46]. The study included: (1) a lab-based evaluation of different sound characteristics (direction, identity, loudness) and visual and haptic feedback designs; (2) a contextual design probe at various locations to evaluate the relevancy of feedback in real-world contexts; and (3) semi-structured interviews exploring the overall experience, feedback preferences, contextual factors, filtering criteria, and privacy concerns related to smartwatch-based sound awareness.

Participants rated the sound's identity as the most useful of the three sound characteristics presented during the study, reiterating DHH users' interest in robust sound recognition tools for accessibility from prior work [13, 37]. Participants preferred designs combining visual and haptic sound feedback, with a simple vibration being preferred in most cases. The study also highlighted the need for adjustable sound filtering, as participants identified diverse use cases for sound awareness in different environments and expressed concerns about universal filtering methods. Overall, these findings emphasized the importance of sound recognition tools that can be personalized to diverse environments and user-specific needs, and they offered valuable insights for developing effective sound feedback on smartwatches for DHH users.

Investigating Audio Data Collection. The second study, detailed in Chapter 4, explored how DHH users capture, interpret, and conceptualize audio data for automatic sound recognition (RQ2). Following my evaluation of contextual sound feedback on a smartwatch, I collaborated on Jain *et al.*'s *SoundWatch* [72], a functional smartwatch-based sound recognition application using a pre-trained model for 20 sounds of common interest to DHH people. DHH participants in a user evaluation appreciated *SoundWatch*'s filtering options but desired to add custom sound categories (*e.g.*, footsteps) in addition to those available in the pre-trained model—reiterating requests from in prior work [71]. However, prior work has not explored how DHH users record and engage with audio data for sound recognition models in depth—despite the quality of a training dataset predicating the effectiveness of a sound recognizer.

To fill this research gap, I recruited 14 DHH participants for a field study to understand their experience recording audio in real-world environments [47]. During the initial session, participants completed an interactive machine learning tutorial and discussed non-auditory feedback through waveform and spectrogram visualizations. They then engaged in a week-long field study, recording sounds of interest using a smartphone application with a waveform visualization. Follow-up interviews explored their experiences, strategies, and challenges during recording, their assessment of recorded audio as training data, and their ideas for features to aid with sample collection.

Participants reported positive experiences recording sounds, but they faced challenges with capturing spontaneous, invisible, or hard-to-reproduce sounds. The waveform visualization was crucial for monitoring ongoing recordings, though participants desired additional information to monitor ambient soundscape activity. When considering the recordings in terms of training a sound recognition model, participants' limited frame of auditory reference led to uncertainty toward their recordings' replication of real-world sounds and the location of potential decision boundaries for the model—leading to requests for features that could provide greater insight into the dataset. These findings provide valuable insights into the needs and challenges of DHH users when capturing audio training data and provide design opportunities for the user interface of an accessible personalizable sound recognition tool.

Evaluating an Accessible Sound Recognition Pipeline. The third study, detailed in Chapter 5, explored feedback mechanisms and UI elements to support DHH users with training (RQ3) and the effect of a complete training cycle on their understanding and assessment of sound recognition models (RQ4). Following my field study of DHH users' experience capturing audio data, I collaborated on Jain *et al.*'s *ProtoSound* [67], a functional pipeline for DHH users to personalize their sound recognition models. ProtoSound included several optimizations to support users' needs, including accommodating varying contexts of use, open-set classification, and an example library for difficult-to-produce sounds. However, as it primarily focused on technical improvements, it did not include accessible representations of audio data and was not evaluated with DHH users.

To understand the needs of DHH users throughout an interactive training workflow and effective UI mechanisms to support this process, I designed and built SPECTRA, a prototype for the accessible creation of personalized sound recognition models [48]. I recruited 12 DHH participants to evaluate SPECTRA for sounds in their homes; following a brief tutorial and interview about their expectations for the system, they trained a personalized model for six sounds. Participants used SPECTRA to capture varied recordings of these sounds to create an initial dataset, then selected suitable examples for a training dataset with support from SPECTRA's interactive clustering interface. After training, they reproduced the sounds in a practical assessment of the model, with the option to return to the previous steps if desired. A follow-up interview discussed the positives and negatives of SPECTRA's UI, participants' satisfaction with the finished model, and suggestions for improvement.

Findings highlighted the value of spectrograms and waveforms, the potential of interactive clustering for exploring an audio dataset, and the use of rich text annotations to prompt reflection on sound replication. The study also revealed the impact of hands-on engagement with machine learning on participants' perceptions, tolerance for errors, and mitigation strategies. The results of this work provide guidance for the design of future personalizable sound recognition tools for DHH users, either to create fully custom models or extend existing ones.

1.3 Summary of Contributions

As a whole, my dissertation contributes a comprehensive, empirical understanding of DHH individuals' preferences and needs around personalization in sound awareness tools, including:

1. The utility of different forms of sound feedback for DHH users and how contextual factors can modulate the relevance of that feedback (Ch. 3);

2. The practical considerations and sense-making strategies that DHH people use in recording and interpreting real-world audio data to train a sound recognition model (Ch. 4);

3. A deeper understanding of DHH peoples' training strategies and conceptualization of ML when creating a sound recognition model (Ch. 5).

In addition, my dissertation contributes guidance for the design of personalizable sound awareness technology, including:

4. Characterization of the complementary roles of visual and vibrational feedback in sound awareness devices (Ch. 3);

5. Implications and considerations for designing specialized recording tools to aid DHH users in capturing an audio dataset (Ch. 4);

6. An end-to-end prototype pipeline to assist DHH users with data collecting, training, and practical assessment of a sound recognition model (Ch. 5);

7. Recommendations for UI elements that can facilitate DHH users in interpreting audio data, and training and evaluating a sound recognition model (Ch. 5).

1.4 Authorship Statement

I am the principal author of the research detailed in this dissertation, but it was conducted in close collaboration with my advisor, Leah Findlater, and my colleagues, including Jon Froehlich, Dhruv Jain, Emma McDonnell, Ping Liu, Rose Guttman, and Susanne Kirchner-Adelhardt at the University of Washington. The following chapters are written in the first-person plural to recognize my collaborators' contributions.

Prior publications from my colleagues have also closely informed my work: my advisor Leah Findlater led a large-scale survey of DHH users' sound feedback preferences [37], and my colleague Dhruv Jain led formative studies on speech and sound awareness technology [65, 66, 70] and later, evaluations of prototype sound recognition tools [67, 71, 72]. This work is referenced throughout my dissertation and differentiated from my own where applicable.

Chapter 2

Background & Related Work

To situate my work, I first outline perspectives on disability and the DHH experience that shape my work, followed by prior research on sound awareness needs and preferences among DHH users, sound awareness technologies (focusing on those for sound recognition), and human-centered machine learning.

2.1 Disability Models and DHH Perspectives

Historically, the field of human-computer interaction has focused on addressing accessibility barriers through an impairment-focused lens, often overlooking the broader sociocultural context of disability [95]. This approach aligns with the medical model of disability, treating differences in ability as a deficit that needs to be fixed or cured through medical or technological intervention (*e.g.*, hearing aids, cochlear implants) [43]. In contrast, the social model of disability shifts focus away from the individual, focusing instead on solving societal and environmental barriers—such as by challenging exclusionary norms and advocating for greater accommodation (*e.g.*, captioning, sign language interpreters) [33, 124]. As an extension of the social model, the cultural or diversity model views disability as a distinct part of one's sociopolitical identity to be valued and celebrated [5]. While a DHH individual may choose to be viewed within any of these models at different times [19], this dissertation primarily draws from the social model, considering how sound awareness technology can address environmental barriers to access and promote inclusion. My work follows Mankoff *et al.*'s call for accessibility researchers to consider target populations as experts in design and move beyond technological expectations determined by society [95].

My work also acknowledges the value of the cultural model in recognizing and embracing Deafhood, the process by which DHH people come to actualize their identity [85, 86]. An individual with hearing loss may choose to identify as Deaf (capital 'D'), deaf, or hard of hearing [84, 105]. Individuals who identify as Deaf follow an established set of shared norms, behaviors, and language [86], which contrasts with hard of hearing or deaf individuals, for whom deafness is primarily an audiological experience. In particular, these perspectives motivate my work's focus on personalization, recognizing that users' needs and preferences are shaped by a complex interplay of social, cultural, and individual factors. Bauman & Murray's conceptualization of *Deaf gain* [11] further highlights the unique sensory orientation and use of visual-spatial language within Deaf culture, framing these experiences as a source of strength rather than deficiency. My work adopts the idea of Deaf gain by drawing on DHH expertise in designing visual technology that supports their diverse ways of engaging with the world, rather than simply replicating hearing-based experiences.

Prior work has consistently shown widespread interest in sound awareness technology among DHH people [13, 37, 70, 98], but this interest is modulated by cultural factors. People who prefer sign language are generally less interested in sound awareness than those preferring oral communication [13, 37]. In a Deaf cultural context, deafness is not a disabling phenomenon, and *audism*—discrimination based on hearing ability [61]—is absent. Many within the Deaf community oppose invasive technological interventions (*e.g.*, cochlear implants [11, 138]), opting for built environments that maximize visual-spatial access and minimize reliance on sound information (DeafSpace [31]). Many Deaf individuals use assistive technology to engage with the hearing world (*e.g.*, closed captioning), but there is potential for this technology to perpetuate audism by imposing normative hearing standards. Assistive technology can be "double-edged" [108]: while it may offer a tangible redress of sociocultural disadvantage, it can also reinforce notions of disability as solely an individual problem solvable with the correct prescriptive device [92, 125]. My work does not aim to replace or substitute hearing ability but rather to develop sound awareness technology that complements the existing tools and strategies used by DHH individuals to navigate the world.

2.2 Sound Awareness Needs and Preferences

Prior work reveals a range of sound awareness needs and preferences among DHH people, influenced by cultural identity, social context, and physical location. When discussing sounds of interest, DHH users generally rank awareness of safety and urgent sounds (*e.g.*, alarms, sirens) as most important, followed by sounds that indicate others' presence (*e.g.*, footsteps, door knocks) and appliance alerts (*e.g.*, oven timers, pop-up toasters) [13, 37, 70, 98, 103, 104, 127]. However, this preference can depend on one's cultural identity; Findlater *et al.*'s survey of 201 DHH individuals found that hard of hearing users may be more interested in certain sounds (*e.g.*, phone ringing, spoken conversations) than those identifying as Deaf or deaf [37]. The specific sounds of interest to a DHH individual can also be highly personal, including babies and children [98], name calls [13], and specific sounds from the home [71].

Sound awareness needs and preferences are also strongly influenced by contextual factors. DHH respondents in Findlater *et al.*'s survey [37] predicted that social context (*e.g.*, with friends *vs.* strangers) would impact their use of a sound awareness tool, and a majority desired to have sound filtering rather than being informed of all sensed sounds. Sound awareness needs may additionally differ by physical location, with past work asking interview or survey respondents about the relevance of sound information at home, work, or while mobile [13, 98]. In contrast to studies that hypothetically asked about varied contexts of use [13, 37, 98], participants in my evaluation of smartwatch-based sound feedback (Ch. 3, [46]) experienced multiple real-world settings.[1]

Prior work has also studied the characteristics of sound desired by DHH individuals, finding that some (identity, location, urgency) are generally viewed as more important than others (volume, duration, pitch) [13, 37]. However, the relative utility of this information may also differ by location or how the information is conveyed. For example, in the home, sound identity and location may be adequate [71], while directional indicators are important when mobile [103].

My dissertation is informed by the above work, and I contribute to this literature by exploring users' preferences for sound feedback and filtering options across different physical locations. Further, the rang-

[1] This work came before Jain *et al.*'s SoundWatch [72], which presented sound filtering across different contexts using a functional sound recognition tool.

ing sound awareness needs and preferences among DHH individuals motivate my research's focus on personalizable tools.

2.3 Sound Awareness Formats and Feedback

Sound awareness technologies can take stationary, handheld (*e.g.*, smartphone or PDA), or wearable formats. Early HCI research focused on stationary designs such as desktop displays [56, 98, 99, 137]. More recently, however, attention shifted toward portability, including smartphone apps for environmental sounds [13, 103, 127] or automatic captioning—features that are now available on Android [50] and iOS [6] platforms. Wearable solutions have also emerged, with head-mounted [52, 66, 80, 109] and wrist-worn wearable devices [72, 76, 102, 104] offering alternative options for sound awareness.

While these portable tools show promise, user evaluations—when present—have been limited to the lab or a single environment (*e.g.*, a classroom setting [127]). These studies have highlighted the tools' potential to provide communication support [66, 76, 109] and alert to urgent situations [13, 103, 104]. However, at the time of my research,[2] prior work had not probed the practical issues of how to manage soundscape complexity within sound feedback and sampling tools—which are key areas of study within this dissertation (Ch. 3 [46]& Ch. 4 [47]).

For sound awareness feedback, several studies recommend combining visual and vibrational information [13, 37, 84, 102], and the ability to do so is seen as a strength of smartwatches [37, 70, 102, 104]. User evaluations of prototypes that combine visual and haptic feedback, however, have been limited to using vibration as a secondary modality to draw attention to the visual information [13, 102, 103, 104, 127], and have not compared alternative approaches for combining the two modalities. In contrast, my work explores combinations of visual and vibration feedback, including using tactons [16] to convey richer feedback via vibration.

As tactile approaches provide much lower information throughput than visual approaches [32], the extent of sound information to be conveyed haptically and its combination with visual feedback is an open research

[2] Since my work's publication, more evaluations of portable sound awareness tools emerged: Jain *et al.* explored sound filtering in multiple contexts with a functional smartwatch prototype [72], and Huang *et al.* conducted a three-week field study with a similar prototype used in daily life [60].

question. For example, Jain *et al.* explored sound accessibility in virtual reality (VR) contexts [68, 90] and characterized a variety of sound-to-visual/haptic mappings in existing VR apps and games [69]. Wearable vibrotactile approaches without visual displays have also been studied, often for sensory substitution of auditory information [23, 42, 64, 133, 149, 151]. Obtrusive form factors were impractical for everyday use (*e.g.*, waist-mounted [23], neck-worn [42]), but early wrist-worn vibrotactile sound aids showed more promise [42, 133]). However, frequent vibrational feedback has a high attentional cost, especially in noisy environments [64, 111]—a concern explored further in my work (Ch. 3).

The emergence of smartwatches as mainstream devices [37, 102, 103] offers a unique opportunity to merge visual and haptic sound feedback without the stigmatization of dedicated assistive devices [125]. Compared to smartphones or head-mounted displays, smartwatches are the preferred portable sound awareness format among DHH individuals [37]. Smartwatch interactions typically consist of only brief glances [114], so visual designs need to emphasize glanceability and space efficiency [12, 21]. Smartwatch-based haptics are typically employed for simple notifications: the watch vibrates, and the user can either ignore the alert or check the watch screen for detail [114]. However, the wrist has high perceptual sensitivity to vibrotactile patterns, opening possibilities for more complex haptic output [89]—such as for video games [110], passive Morse code learning [123], or sound awareness, as explored in my work.

2.4 Sound Recognition Tools

Information about a sound's identity is commonly desired among DHH users [37, 98], and surveys of DHH individuals reveal a widespread desire for sound recognition tools that can notify when a sound is detected [13, 37, 67, 72]. An early project by Matthews *et al.* [97] demonstrated the potential value of such tools, exploring human-powered transcription of both speech and non-speech sounds from recordings captured via PDA, despite misjudging relevant sounds and limitations in scalability and users' privacy.

Following advancements in digital signal processing and machine learning, more recent work has aimed to provide broad sound recognition support by employing pre-trained classification models [71, 72, 102, 127]. These models, while offering a broad range of recognizable sound categories, often fall short in addressing the unique needs and environments of DHH users [60, 67] For example, Liu *et al.*'s deployment of a smartphone app within a school setting revealed common concerns with the accuracy of brief sound events [127], while Jain *et al.*'s in-home deployment of a tablet-based model revealed inconsistent classifications and requests to accommodate specific sound preferences [71]. Huang *et al.* [60]'s field study of a smartwatch-based sound recognition app (conducted after my work in Ch. 3 [46]) found more extensive issues with background noises, variability of real-world sound events, and misclassification of similar sounds. These findings suggest models need greater personalization options for users' unique environments and sounds of interest.

DHH users frequently request personalization options for as-needed support of their individual needs (*e.g.*, [37, 98, 103]). Some projects (following my work in Ch. 3 [46]) explored options for DHH users to filter notifications for certain sounds [71, 72]—a customization option since added to Apple and Android smartphone platforms—but they stopped short of adding or modifying sound classes through user-provided recordings. AdaptiveSound [28] explored user-driven fine-tuning of models through corrective feedback, but the system did not support custom sound categories and offered little insight into their model's shifting decision boundaries.

Deeper explorations of user-driven personalization have been limited. Bragg *et al.* [13] conducted a Wizard-of-Oz study of a smartphone app for training custom models, but this study lacked a functional model and involved recording only two sounds (alarm clock, door knock) in an office setting—an experience that does not represent the varied use cases, sounds, and environmental noise of daily life. In contrast, my field study in Ch. 4 [47] details the experience of DHH users recording chosen sounds in everyday environments, providing insights into the practical challenges of this task.

Building on this, Jain *et al.* [67] surveyed 472 DHH users of Android's sound recognition feature and developed ProtoSound, a training pipeline with technical considerations for DHH users (*e.g.*, limited data, contextual flexibility). However, ProtoSound was not evaluated with DHH users and did not include a front-end interface to visualize audio data [67], leaving a gap in understanding how such a system would be used and experienced in practice. This gap motivated my development and evaluation of SPECTRA, a personalizable sound recognition prototype detailed in Ch. 5 [48].

2.5 Human-Centered ML for Accessibility

Human-centered machine learning aims to design and build automated systems that can fulfill user goals, fit user-specific contexts, and accommodate people without programming experience [36, 118]. This work is particularly valuable for the field of accessibility, where data-driven assistive technologies may be improved through personalization for an individual's needs [74]. However, training an ML-enabled application as a personal assistive technology can itself be inaccessible when it requires skills and abilities similar to those the application is intended to support [38, 106]. For example, a blind or visually impaired user is likely unable to use visual feedback when capturing images for personalizing an object recognizer—a challenge that Kacorri *et al.* and others (*e.g.*, [75, 129]) first examined and more recently began addressing through active feedback techniques to assist in image capture [58, 88]. Similarly, my work explores the unique challenges that DHH individuals face when working with audio data for sound recognition and presents opportunities to better support them in this process.

Several human-centered ML paradigms have emerged to support non-expert users in building ML models. *Automated Machine Learning* (AutoML) (*e.g.*, [24, 143]) systems allow novice end-users to provide a large batch of labeled data, while traditional ML tasks—such as feature engineering and model selection—are completed automatically [29, 147]. In contrast to AutoML's black box approach, interactive machine learning (IML) treats end-users as "humans-in-the-loop" who iteratively engage in building and refining ML models [3, 30, 117, 132]. An IML workflow (like SPECTRA's) involves a quick loop between model training, feedback, and usage, where the user may provide indicative samples, describe salient features, or select high-level model parameters [30, 132]. Interactive machine *teaching* [117, 152] takes IML engagement one step further by positioning the human-in-the-loop in the role of the model's teacher, emphasizing human expertise to guide machine learning [142].

Despite the potential of these paradigms for accessibility, they also assume that an end-user has domain expertise and data literacy—an assumption that may not be true for DHH users and audio data. Human-centered ML work with DHH users is limited to a workshop study by Nakao *et al.* [107], which sought to characterize ML understanding among DHH participants through their collaborative use of a trainable sound recognition interface. Participants were uncertain about the contents and quality of sound data due to the absence of non-auditory feedback (*e.g.*, visualizations) within the system. This is a critical gap, as the quality of training data directly impacts the generalizability of an ML model, and dataset refinement is an invaluable strategy for performance optimization [57]. My work investigates non-auditory data representations to improve the data literacy of DHH users throughout the stages of an interactive training process.

Outside of accessibility, IML research for audio has primarily focused on sample annotation and labeling (*e.g.*, [63, 78, 79, 126]). For interactive sound recognition, Ishibashi *et al.* [63] explored visualization options (*e.g.*, spectrograms, thumbnails) for browsing large sets of unlabelled audio samples via a clustering interface. Google's Teachable Machine [17] allows non-expert users to quickly train a personal sound recognition model with their own audio examples, but it provides limited audio visualization (low-resolution spectrograms) and lacks information on the quality of a user's training set (*e.g.*, clustering feedback). Nakao *et al.*'s work [107] explored a comparable workflow (without visualizations) with non-expert DHH users, allowing them to create training sets from their recordings or choose recordings from a sound library. After a shared hands-on experience, DHH participants identified additional use cases and showed an improved understanding of ML; however, some found it challenging to review samples and define classes for sounds they were familiar with but unable to hear.

This prior work—in combination with others [13, 67]—begins to outline an interface design space for personalizable sound recognizers for non-expert DHH users. As a next step, we built and evaluated a specialized IML workflow tailored to the unique needs of DHH users, advancing understanding of how non-auditory data representations can support this population during interactive sound recognition tasks and yielding insights for designing future tools in this area.

My work contributes to this body of work by exploring how DHH users collect, interpret, and evaluate audio data for personalized sound recognition (Ch. 4 [47]), along with developing and evaluating an accessible prototype with non-auditory feedback to support DHH users when interactively training sound recognition models (Ch. 5 [48]).

Chapter 3

Evaluating Sound Feedback Across Contexts[1]

3.1 Introduction

Advances in wearables and audio processing provide new opportunities for portable sound awareness solutions [45, 96, 109]. A survey of 201 Deaf and hard of hearing (DHH) participants [37] found that smartwatches were the most preferred portable device for non-speech sound awareness compared to smartphones and head-mounted displays. Further, smartwatches were seen as useful, socially acceptable, glanceable (for all sound scenarios except captions), and advantageous because they could provide both haptic and visual feedback.

Most prior work, however, has focused on smartphones and older handheld devices [13, 97, 127], head-mounted displays [52, 66, 109], and custom wearable systems that provide limited information through a single modality (*e.g.*, [76, 149]). For smartwatches specifically, Mielke and Brück [102, 104] conducted a preliminary lab study with six DHH participants using a Wizard of Oz interface; a simple vibration occurred

[1] This chapter includes materials originally published in [46], which explored user preferences for smartwatch-based sound feedback in different contexts.

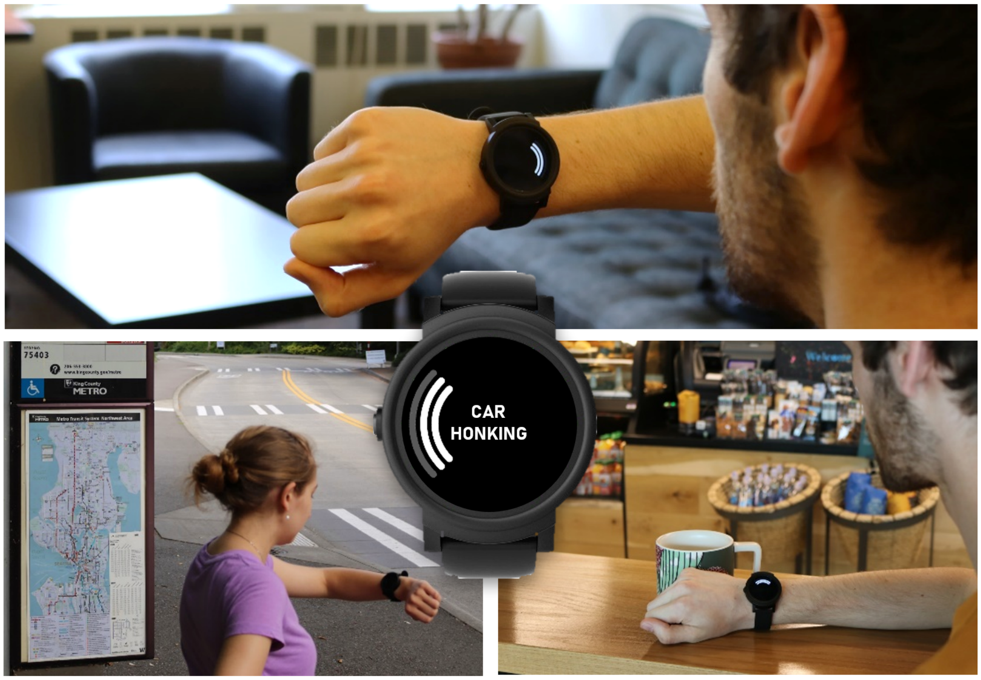


Figure 3.1: Participants used a smartwatch for sound awareness in three contexts during an *in-situ* exploration: a student lounge (top), bus stop (left), and café (right).

when the wizard triggered feedback, and a visual sound was displayed. User reactions were generally positive, but given the limited nature of that study, many design questions remain. For example, how should a design most effectively combine visual and haptic feedback on the smartwatch? And, is there a role for haptic feedback that is more complex than simple vibration?

Further, smartwatches and other portable sound feedback systems will need to function in complex soundscapes. Constant vibrational sound notifications are not desirable [111], for example, and some projects have examined filtering sounds based on identity [13, 98] or loudness [133]. Beyond these initial steps, there has been little investigation into how to design sound feedback for complex soundscapes. How should sound filtering be designed, and what are the implications for filtering when both visual and haptic feedback modalities are present?

To address these questions, we conducted a three-part study with 16 DHH participants: (1) a Wizard-of-Oz evaluation of a smartwatch prototype comparing three designs that offer visual feedback plus different levels of vibrational feedback (none, simple, tacton) and exploring haptic and visual techniques to portray three sound characteristics (*direction*, *identity*, and *loudness*); (2) an *in situ* experience (Figure 3.1) where participants visited three locations (café, bus stop, student lounge) and used the watch to receive a preset sequence of sounds typical of the location; (3) a semi-structured interview covering the user's overall experience, the feedback design, sound filtering options, and possible privacy and social acceptability issues.

Our findings confirm the importance of combining visual and haptic feedback for sound awareness [37, 102, 104], but extend past work by showing that vibration is particularly important for push notifications to draw attention to visual details and that haptics have the potential to support more discreet and immediate sound alerting. Participants saw utility in vibration patterns (tactons), emphasizing this as a promising direction for future work. In terms of soundscape complexity, the *in situ* experiences caused all participants to request sound filtering—particularly to limit haptic feedback—with varied advantages seen for filtering by sound *identity*, *loudness*, or *direction*. We also report on important concerns that must be addressed in future designs, including learnability of the tactons, the possibility of distraction, and issues of trust in the system.

This chapter contributes: (1) a deeper understanding of the complementary roles of visual and vibrational feedback for a wearable sound awareness device; (2) evidence of the potential for small sets of haptic patterns to convey sound information; and (3) characterization of initial subjective responses to soundscape complexity and potential means of managing that complexity based on three pre-set locations. We also close with a discussion of design considerations and directions for future work.

3.2 Method

We employed a design probe method with 16 DHH participants to elicit user preferences for watch-based sound awareness. The method included a Wizard-of-Oz prototype evaluation in the lab, a demonstration of how such a system could work in practice in three *in situ* settings (*e.g.*, in a café), and a semi-structured interview. We investigated haptic feedback preferences, sound filtering, contextual factors, privacy, and social concerns.

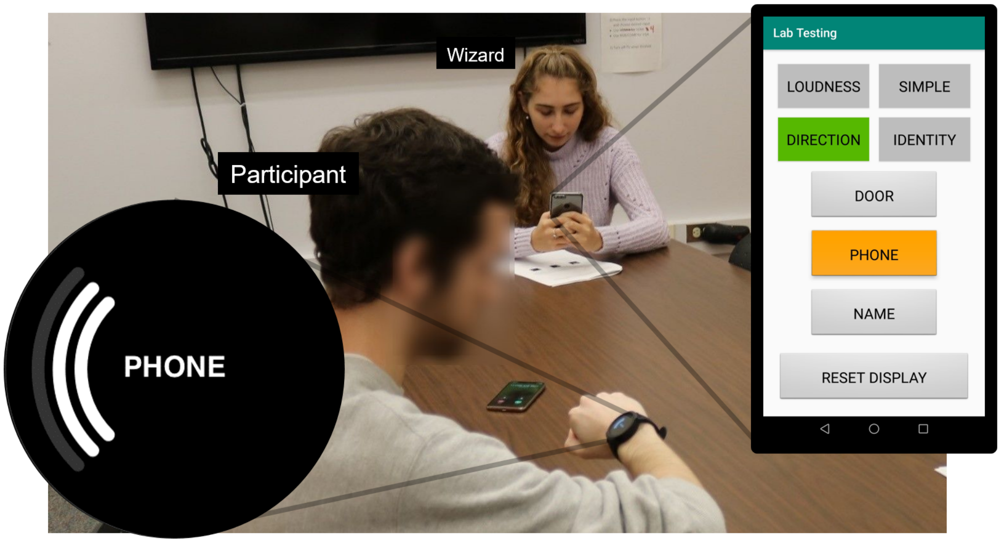


Figure 3.2: The Wizard-of-Oz prototype used for our lab evaluation. The wizard used a smartphone app (right) to remotely trigger visual (left) and vibrational feedback after sound events, such as a phone ringing (center).

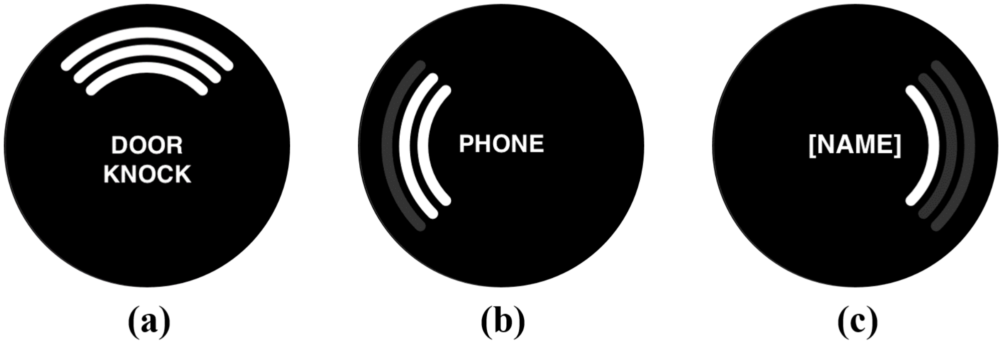


Figure 3.3: The smartwatch display shows the *direction*, *loudness*, and *identity* of sounds, such as: (a) a loud door knock in front of the wearer, (b) a moderate phone ringing to the left, and (c) a quiet name called to the right.

3.2.1 Smartwatch Prototype

To provide a realistic user experience and to enable us to compare different types of visual and vibrational feedback, we designed a Wizard-of-Oz prototype that consisted of two parts: a "wizard" interface running on an Android-based smartphone (Honor 7X) and a participant interface running on an Android-based smartwatch (Mobvoi Ticwatch E). The wizard interface could trigger events on the watch via Bluetooth (Figure 3.2).

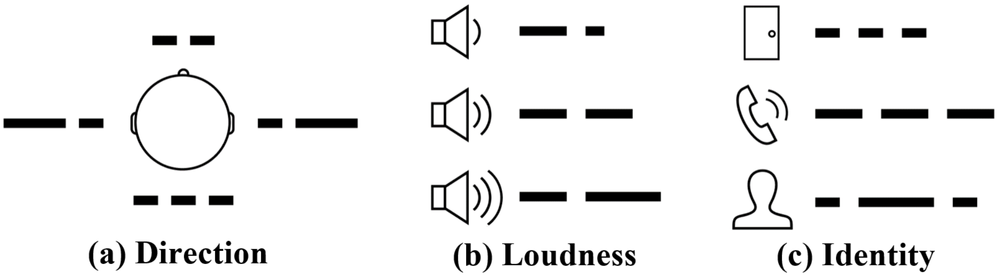


Figure 3.4: Visual illustrations used to introduce participants to the three tacton sets: *direction*, *direction*, and *identity*. Lines of varying length indicate the relative duration of each vibration within a tacton. Tactons for four *directions* (a) were based on PocketNavigator [112], while three tactons for *loudness* (b: low, medium, high) and *identity* (c: door knock, phone ring, name call) were our own design.

Visual feedback

Informed by [12, 99], we designed our visual feedback with a minimalist, high-contrast, and glanceable aesthetic (Figure 3.3). The display conveys three properties of sound: *direction* as three 90° arcs pointed towards the sound source; *loudness*, which fills the directional arcs depending on sound amplitude (three discretized levels), and *identity*, which shows the classified sound event as text in the screen's center. For this prototype, we implemented direction relative to the wearer's torso (in front, behind, to the right, or to the left), assumed perfect sound classification, and did not support co-occurring sounds (only the loudest sound was shown). We return to these design decisions in this chapter's Discussion (Section 3.4).

Haptic feedback

Past work has paired visual feedback with a simple vibration for notification [13, 102, 103, 104, 127]. In addition, we explore vibrational patterns (tactons), including how to best convey sound characteristics with tactons. Based on the capability of our off-the-shelf smartwatch (*e.g.*, vibration output at fixed frequency and amplitude), we designed our haptic feedback as follows:

Simple vibration: A single 500ms vibration occurs with each sound event to notify the wearer.

Tacton sets (vibration patterns): Informed by Brewster and Brown's study of tactile icons ("tactons") [16], we separate small sets of tactons to convey each of the following sound characteristics: *direction*, *loudness*, and *identity*. Each tacton consisted of a pattern of on/off vibrations at a constant intensity between 200–1200ms

Table 3.1: Demographics of study participants. HH = hard of hearing.

ID	Age	Gender	Cultural Identity	Self-reported Hearing Loss
P1	19	NB	deaf	Profound
P2	62	M	deaf	Profound
P3	53	M	deaf	Profound
P4	54	W	deaf	Profound
P5	33	W	Deaf	Profound
P6	46	M	Deaf	Severe
P7	51	W	Deaf	Profound
P8	56	M	deaf	Severe
P9	61	M	deaf	Severe
P10	61	M	HH	Moderate
P11	69	W	HH	Moderately severe
P12	86	M	HH	Moderately severe
P13	74	W	Deaf and HH	Profound
P14	69	W	Deaf	Profound
P15	69	M	Deaf	Profound
P16	27	W	Deaf	Profound

long. Prior work shows DHH users prefer patterns over sustained vibration for attention-getting [53], and temporal patterns at a constant intensity level are easier to discern than those with varied intensity [89]. For *direction*, we defined tactons for left, right, front, and behind based on Pielot *et al.*'s PocketNavigator [112], a tactile compass on a smartphone designed using the tactons framework [16]. We then created two more tacton sets of three patterns each: one set for *loudness* and one for *identity* (Figure 3.4); for example, a short-short-short vibration pattern indicated a door knock, while a short-long-short pattern indicated a name call.

3.2.2 Participants

We recruited 16 DHH participants through direct email, a hearing loss organization, and snowball sampling. Eight participants identified as men, seven as women, and one as non-binary. Participants were 55.6 years old on average (*SD*=17.7, range=19–84). Fourteen participants reported using hearing devices: eleven used hearing aids, five used cochlear implants, and two used both (Table 3.1).

Procedure

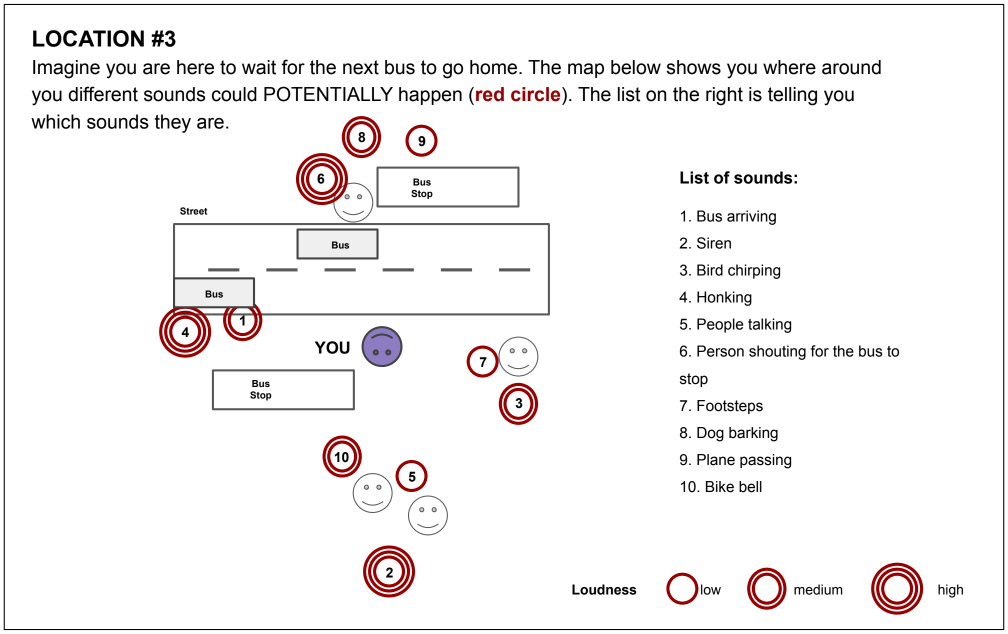
Before the study session, participants completed an online questionnaire to collect demographics, current use of sound awareness technologies, important sounds in daily life, and initial reactions to new sound awareness solutions. The in-person procedure took place on a university campus and lasted 90 minutes. Sessions were led by the first author, with one of three rotating research team members acting as a wizard. We allowed participants to request communication support for the session: six opted for a sign language interpreter, and two opted for a real-time captioner. Instructions and interview questions were presented visually on an iPad, while responses and follow-up discussions were spoken and translated to/from ASL.

The session consisted of three parts, the first and third of which took place in a quiet conference room. P11 had to leave unexpectedly after the lab-based design probe (Part 1) but returned to complete the remainder of the protocol 12 days later.

Part 1: Lab-based design probe (30 min): To give participants an idea of how a smartwatch-based sound awareness system could sense and convey different sounds, we presented a Wizard-of-Oz prototype in a lab setting. The participant sat at a conference table facing the door, with the facilitator on the opposite side and the wizard to the participant's left (Figure 3.2). After discussing current sound support strategies and soliciting reactions to smartwatch-based sound awareness, the participant placed the smartwatch on their preferred wrist (in contact with skin).

To gradually familiarize the participant with our Wizard-of-Oz prototype, we introduced three feedback designs in order of increasing complexity: *visual only*, *visual+simple vibration*, and *visual+tacton*. For the first two designs, we included a short description and three example sound events with different *direction*, *identity*, and *loudness* combinations (*e.g.*, Figure 3.4): (1) door knock, performed three times at high volume on a door in front of the participant; (2) phone ring, played at a moderate volume to the left of the participant; and (3) name call, spoken at low volume while standing to the participant's right. For each sound, the wizard remotely triggered the appropriate feedback on the watch.

For the third design (*visual+tacton*), we presented the three tacton sets (direction, identity, loudness) in counterbalanced order, with participants randomly assigned to order (because there were 16 participants, two of the six orders only had two participants) For each tacton set, participants were given a visual reference sheet depicting the vibration patterns for the tactons (*e.g.*, Figure 3.4), as well as a demonstration of what each tacton felt like without any visual feedback. For this demonstration, the facilitator clapped at



LOCATION #3

Imagine you are here to wait for the next bus to go home. The map below shows you where around you different sounds could POTENTIALLY happen (**red circle**). The list on the right is telling you which sounds they are.

List of sounds:

1. Bus arriving
2. Siren
3. Bird chirping
4. Honking
5. People talking
6. Person shouting for the bus to stop
7. Footsteps
8. Dog barking
9. Plane passing
10. Bike bell

Figure 3.5: Sound map used for the bus stop location in the contextual probe. These maps oriented participants to the physical space and prepared them for the ten preset sounds on the smartwatch. The facilitator guided participants to each location, asked them to face in the direction indicated by the purple face while pointing out physical landmarks, and then initiated the smartwatch feedback.

three different volumes in front of the participant (*loudness*), clapped at the same volume in four locations around the participant (*direction*), and created door knock, phone ring, and name-calling sounds in front of the participant at constant volume (*identity*). Finally, the facilitator made the three example sounds described earlier (knock, ring, and name), with the wizard triggering both visual feedback and the appropriate action.

After being introduced to all three feedback designs (*visual only*, *visual+simple vibration*, *visual+tacton*), participants (1) rated the utility (*i.e.*, "useful in everyday life") of the three sound characteristics displayed on the watch in a set order (*direction*, *identity*, *loudness*); (2) rated the utility of each feedback design in random, counterbalanced order to minimize the order effect from their demonstration; and (3) discussed each tacton set in the same order they were presented. Note: Because the study focused on the participant's subjective experience, we did not specify a hypothesis for this rating data, and all quantitative results were considered secondary to the interview responses.

Part 2: Contextual design probe (25 min): To probe participants' responses to the effects of context on sound awareness, we visited three campus locations (student lounge, café, bus stop) and presented a preset sound scene at each one. Locations were visited in a set order, following this scenario:

> *Imagine you are on your way home but forgot your water bottle in the student lounge upstairs. After you pick it up, you go to [the café] in the building next door to pick up some coffee and then go to the bus stop to catch the next bus home.*

In each location, participants were shown a map on the iPad to orient themselves to the preset sound scene (*e.g.*, bus stop map in Figure 3.5). Each map included ten numbered sounds typical of the area, with circles around the number to indicate loudness. After the participant had reviewed the map, the wizard triggered the watch to display the list of sounds in sequence, with three-second pauses between sounds. We chose to have the watch visually convey loudness and direction but not identity. We instructed participants to view feedback from the watch and connect it to each potential real-world sound source as a holistic experience to ground discussion after returning to the conference room. Participants were asked to hold in-depth discussions until after the visits.

Additionally, to spur participants to consider different sound filtering options, we employed simple vibration feedback but varied how it worked across the three locations. Instead of having vibrations occur for all ten sounds, which could be overwhelming in practice, the vibration notification only occurred for (1) the top three *loudest* sounds, (2) the three sounds occurring *behind* the participant, or (3) three of the more important sounds *identified* by the watch. For the latter condition, we imagine that a personalized sound system such as that proposed by Bragg *et al.* [13] would support sound feedback by allowing the user to specify a small set of high-priority sounds to identify. The pairing of contextual location (lounge, café, bus stop) with vibration for *loudness*, *direction*, or *identity* was presented in counterbalanced order, with participants randomly assigned to an order.

Part 3: Semi-structured interview (20 min): Finally, we asked semi-structured questions on participants' overall experience with the system, exploring contextual factors, filtering options, social acceptability, and privacy issues surrounding smartwatch-based sound awareness.

Data and Analysis

Session transcripts were analyzed using an iterative coding approach [14]. Two researchers independently read the first four transcripts and identified a small set of potential codes to form high-level themes. The researchers then met and developed a mutually agreeable codebook with a two-level hierarchy to apply holistically to the data. As additional transcripts came in, the two researchers split the data by odd and even participant numbers and independently coded every other transcript. Upon receipt of the final two transcripts (P15 and P16), the two researchers agreed we had reached thematic saturation. Both researchers then randomly coded one of the other's transcripts to check for inter-rater reliability. Codes with low Cohen's kappa scores were used to identify areas of disagreement and were updated, merged, or removed from the codebook until 71 codes remained (10 first-level, 61 second-level). Each researcher applied the updated codebook to the other eight transcripts they had not yet analyzed. Following another inter-rater assessment, the codebook was considered final, with an average kappa of 0.72 ($SD = 0.14$) and raw agreement of 0.91 ($SD = 0.07$). Again, all code applications were reviewed, and disagreements were resolved through consensus.

For ordinal rating scale data, we used R to perform Friedman tests, a non-parametric alternative to repeated measures of one-way ANOVAs. In cases of significance, Wilcoxon signed-rank tests with a Bonferroni correction were used for post-hoc pairwise comparisons.

3.3 Findings

Participants reported using their hearing aids and/or cochlear implants, as well as smartphones for sound awareness—the latter primarily for automatic captioning ($N = 8$), such as via Android's Live Transcribe. When discussing hearing aids and cochlear implants, participants described well-known limitations [27], including discriminating between sounds ($N = 10$), inadequate background filtering (10), and poor speech comprehension (6). Sounds of interest reflected past work (*e.g.*, [13, 37, 99]): social sounds and alerts were important, while indoor sounds (*e.g.*, doors, typing) and background noise (*e.g.*, birds, traffic) were less desired.

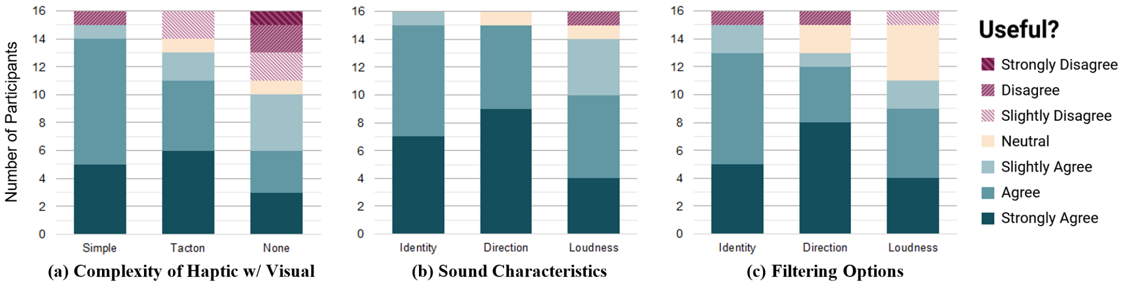


Figure 3.6: Utility of (a) haptic complexity, (b) sound characteristics, and (c) filtering options, with conditions ordered by usefulness.

Before presenting our prototype, we asked participants to share their thoughts on using a wearable device for sound awareness. Similar to findings by Findlater *et al.* [37], most participants responded positively: nine were very or extremely interested, while the remainder were only somewhat ($N = 5$) or slightly (2) interested. All but one participant (P14) saw potential in using a smartwatch for this purpose, although 10 participants also predicted limitations, most notably the small screen size. Below, we cover findings from the in-lab comparison of our three feedback designs, discuss themes that emerged through the *in situ* experiences, and synthesize topics overlapping both parts of the study.

3.3.1 In-Lab Comparison

Participants evaluated three feedback designs on the smartwatch: *visual only*, *visual+simple vibration*, and *visual+tacton*; the latter design was separated into three tacton sets to convey sound *identity*, *direction*, and *loudness*. We report reactions to the three designs, followed by confirmatory findings on general sound feedback preferences.

Reactions to and Preferences for the Three Designs

Overall, participants felt the designs with vibration were more useful than the visual-only design. Perceived utility ratings for all three designs are shown in Figure 3.6 (a). A Friedman test showed that the impact of these design options on utility ratings was significant ($\chi^2_{(2,N=16)} = 6.107$, $p < .05$). No post-hoc pairwise comparisons were significant after a Bonferroni correction, but the qualitative findings below provide insight into the significant main effect.

Visual Alone. Eight participants mentioned unprompted that visual feedback offers higher information throughput compared to vibration feedback and that the visual-only design also offers *"the option to not want to be bothered [by vibration]"* (P13). Thirteen participants were concerned that without vibration, they would miss sounds: *"I'm not going to be looking at my watch every two minutes when I'm out and about"* (P15). Still, across all conditions, participants reaffirmed the importance of visual sound information, such as: *"It's nice to have visual and the sensory input as well [but] I mean without the visual, I feel like there's not really a point."* (P10).

Visual with Simple Vibration. For visual+simple vibration, the key advantage identified by only one participant (P12) was how it could push sound notifications to the wearer. For example, P9 liked that it would support *"alerting to a situation"*, while P5 said it would *"trigger you to look at your watch"*. However, frequent notifications were a concern ($N = 4$), as captured by P16, *"I don't want it to be constantly vibrating because it's a noisy world."* In terms of the specific vibration design, three participants wished it were more prominent, for example, *"Because I think if you're outside moving around or doing something, you might not feel it as much"* (P15). This finding suggests that the vibration intensity should be adjustable.

Visual with tactons. Participants felt that the primary benefit of tactons was minimizing the need to look at the watch face, which eight participants viewed as speeding up their response time. Providing more non-visual detail than a simple vibration would also allow the wearer to *"determine whether it was worth looking at"* (P8) and could provide sound support in a socially acceptable manner:

> *"Let's say I'm at a meeting, and I'm trying to listen to somebody, but I don't want to be rude and look at my watch when they're talking. [...] 'Ah-ha!' Somebody is calling my name, and then I don't have to look [at the watch]."* (P3)

At the same time, important design concerns arose. Some participants ($N = 8$) worried about the effort and time required to use tactons. For example, P13 commented on the difficulty of interpreting tactons while physically active: *"Unless you're just sitting here, it's going to be hard to know what's happening."* Further, P6 commented on the time required to absorb tacton-based information: *"I had to wait and wait, and then the vibration had finished. And by the time I finished decoding what it was, [...] I'd missed whatever happened."*

Participants were also concerned about learning tactons ($N = 9$). P11 described this challenge while also appreciating the possible long-term benefits, saying, *“It might take you a while to learn it, but after you learn it, then it would be automatic.”* As another example, P14 contrasted the difficulty of learning tactons to the ease of the visual feedback: *“It would work after I got used to it, whereas [with] just the visual you wouldn’t need to get used to it.”*

To mitigate learning and recall problems, several participants ($N = 8$) suggested using only a small, simple set of tactons: *“I think three is enough to memorize, but [more] would be hard to distinguish”* (P16). Another approach was to more intuitively match the tacton design to the semantics of the sound ($N = 4$): *“There’s no reason to assume the door knock’s three quick bursts, while the phone is three longer bursts, and a name being called out is this other pattern”* (P2).

Finally, after participants used the three tacton sets (*i.e.*, for sound *identity*, for *loudness*, and for *direction*), we asked which of those three sound characteristics they would most like conveyed via tactons. Responses were split between *identity* ($N = 8$) and *direction* ($N = 6$), with only two participants preferring *loudness*. For example, P10 chose tactons for sound *identity*, saying, *“If it’s a phone call and I’m busy right now, I can ignore it. Whereas, if it’s my wife calling [my name], I better check that out”* (P10). P16, in contrast, felt that *direction* was most immediately actionable, stating: *“When someone calls me [...] I don’t want to have to look at the watch and then look toward the sound.”*

Sound Characteristics in General

As shown in Figure 3.6 (b), most participants felt that all three sound characteristics were useful but to differing extents: a Friedman test showed that there was a significant effect of sound characteristics on utility ratings ($\chi^2_{(2,N=16)} = 8.24, p < .05$). After a Bonferroni correction, no post-hoc pairwise comparisons were significant, but the significant main effect adds evidence to past work showing that sound identity and direction are of greater interest than loudness [37, 98].

Participants also provided open-ended reasoning for their ratings. While participants were generally positive about receiving sound identity information, some ($N = 3$) expressed concerns about accuracy, which also arose later in discussions about sound filtering. For *direction*, three participants felt that this information would be useful because they can identify sounds by hearing, but *"the direction is much harder to pick out"* (P2). Finally, participants mentioned specific ways in which *loudness* could be useful, most commonly related to safety ($N = 5$): *"When something is loud, then it's something that you want to be alerted for"* (P13) and *"[A smoke alarm] is loud for some people, but to me it's not"* (P3).

Summary. Participants preferred to have vibrational and visual feedback rather than visual alone. Both vibrational designs show promise, though tactons may need to be limited to a small, simple set.

3.3.2 Physical Contexts of Use

Following the lab evaluation, participants visited three locations to experience how the prototype might work in context: a student lounge, café, and bus stop. These visits helped participants consider additional aspects of sound feedback, with 11 participants mentioning new use cases and/or increased interest in smartwatch sound awareness. Emergent discussions focused on soundscape complexity and safety, primarily in public or semi-public contexts—but private use in the home also arose.

Soundscape Complexity. After visiting the three locations, participants remarked on the diversity and number of available sounds, as captured by P15:

> *"There's a huge variety of things that you could need to be aware of. [...] You're out there, [and] you don't think about them. But in designing something that could be useful, you do have to think about it. And the complexity of it, of what the environment is, this is eye-opening."*

Despite the challenge of designing for this complexity, busy public locations may be particularly important for sound awareness support. For example, P14 returned from the visits feeling more positive about the watch than she had in the lab:

> *"The [café] is just phenomenal because it's the thing that really gives people anxiety. 'Are they going to hear me? Am I going to hear them?' There's so much ambient noise. In a place like [the student lounge] or in your house with the microwave and whatever, okay, it's quiet. But when you go to a place outside, bus stop, [café], outside your home, this is just... and again in your car, this is just incredible."*

As a result of their *in situ* experiences, four participants changed how they thought about haptic feedback. For P6, the value of the haptic tactons increased: *“There’s just a lot going on, and so if it was a short vibration, I could know to ignore it.”* Similarly, for P14, who went from having no interest in vibration before the visits to saying, *“I realized that even though I do hear the sound, I want that vibration.”* Conversely, the complexity of the soundscape gave P8 and P16 a greater appreciation for the option to *“mute”* (P8) the vibration and only see visual information.

Situational Awareness and Safety. A second set of reactions emerged around situational awareness and safety. All participants except P4 liked the watch for mobile use, with many ($N = 9$) emphasizing personal safety. For example, P16 discussed the utility of sound notifications when walking alone at night: *“I want to know if there’s some sort of noise if someone might be following me.”* Similarly, P3 mentioned using the watch for traffic awareness: *“It could alert you when there’s a hazard, like if I’m riding my bike: ‘There’s a car coming.’”* Five participants mentioned using the watch during outdoor recreation, such as *“thunder”* (P3) and *“a mountain lion”* (P6) while hiking. P6 also imagined the watch could warn of danger in his workplace: (*“[It could] say a shelf of product just fell down”*).

However, not everyone agreed on the watch’s value for situational awareness, with P4 expressing concern that the watch could be distracting and thus reduce safety: *“I would never use something like this to tell me about traffic. Ever. [...] Taking time or being distracted by vibration to look at the watch, it takes me away from my environment.”* Many participants ($N = 8$) also raised the idea of using the watch in a professional or classroom setting to aid in social participation. For example, P16 thought the watch could help her in class: *“Sometimes my teacher will call my name, but I don’t notice that it’s happened, and then I miss the question that’s being directed at me.”* P7 wanted to discriminate *“softer versus louder”* sounds in her work with musicians.

Usage in the Home. While not asked directly, all participants mentioned potential benefits in the home. Almost half ($N = 7$) discussed emergency alerts while sleeping; P4 said, *“To be able to go to bed, put this on, and know if somebody was trying to break down the door or the fire alarm is going off, or maybe the baby is crying.”* Other notable home uses for the watch included awareness of family voices (7) and responding to non-urgent sounds (6), like appliance alerts. For example, P15 recalled once forgetting to turn off his alarm clock, *"and both neighbors on both sides of me in the houses, they checked to see that everything was okay. I was mortified. But if I'd had a watch, it would've said, 'Hey, there's a sound going on,' and that would really have been nice to have."*

Summary. The *in situ* experience with the watch highlighted sound support challenges in busy, public contexts. The watch showed promise for safety while mobile, at home, and for social support in school or professional settings. However, the negative impacts of distraction need to be considered.

3.3.3 Synthesizing Cross-Cutting Themes

Finally, we present cross-cutting themes that emerged across the entire study session, including sound filtering, social contexts of use, privacy concerns, and design suggestions.

How to Deal with Soundscape Complexity: Filtering

Upon returning to the lab, all participants were against conveying every detected sound, reflecting past work [37, 98, 104]. For example, P4 said, *"I don't think I would want [the watch] constantly telling me every sound that came in with directions and arrows,"* while P15 said, *"I think there needs to be some way to filter what you do want to pay attention to, and that's going to differ for everybody."* A few participants ($N = 3$) mentioned wanting visual feedback for all sounds but filtering some vibrational feedback. P7, for example, linked the need for filtering to the ability of hearing people to ignore or attend to sounds:

> *"[Hearing people] have the ability to screen out the sounds because you guys are used to hearing. [...] The vibration, I think, could be a lot for me because it doesn't actually have the ability to filter out the sounds, which is why I prefer to see the visual and I can pay attention to it and then decide."*

Importantly, one participant (P4) was hesitant about adding filtering at all, expressing concern about allowing the device to choose what to filter:

> *"You might be filtering out other awareness that you have built up over years in favor of, 'Well, this thing knows, and in fact, this thing might know better than me, so I'm just gonna ignore my*

> *instinct, I'm not going to bother looking because this will tell me.' [...] I want to hear it all, and I want my own; I want to be able to choose what's more important."*

Filtering as a Function of Sound Characteristic. While prior work has explored notifications for noise above a certain threshold [133] or specific sounds [37, 98], our study is the first to compare several filtering options. At the three *in situ* locations, participants experienced what it would be like to filter vibration notifications based on *loudness*, *direction*, or sound *identity*. Figure 3.6 (c) shows the utility ratings for these three filtering options; while they did not differ significantly, ($\chi^2_{(2,N=16)} = 3.05$, $p > .05$), qualitative comments highlight tradeoffs and possible applications of each.

All but one participant (P4) wanted to filter by sound *identity*, reflecting past work [37, 98]. Many discussed how they would prioritize identified sounds; for example, P2 was excited to see nature sounds: *"I think different people want to know about different sounds. I would like to pick up, like, bird calls, bird chirping."* Three participants, however, were concerned that filtering sounds by type would be technically infeasible, especially in contexts where sounds are unpredictable. P4, for example, said of the café: *"You [the user] can't program in breaking of glass because you wouldn't know that was gonna happen."*

Most participants responded positively to filtering by *direction* ($N = 13$): *"The deaf population, we see things ahead of us and know what's happening"* (P8), but *"things happening behind me, that would be desirable [to know]"* (P13). Safety was often given as the primary reason to filter by *direction* ($N = 5$).

Finally, eleven participants wanted to filter by *loudness*, emphasizing the relationship between volume and important sounds, along with the need to consider different ambient noise levels. P1 was enthusiastic about loudness filtering while recognizing its limitations: *"It might be annoying because there are a lot of loud sounds that aren't necessarily that important, [...] but at the same time, loud sounds are often loud for a reason, so I feel like it's still necessary."* Other participants felt loudness filtering would be useful only in certain contexts; for example, loudness could be useful for P7 in *"a music room"*, while for P5, being notified of loud sounds while at work could be problematic because *"I work close to a fire station."* As another example, P6 experienced loudness filtering at the café and was concerned that the loud ambient noise would create distracting notifications: *"If I'm paying attention to my watch and it keeps vibrating, I might miss my drink come up."*

Summary. Participants requested that sound feedback be filtered, and all three types of filtering (*identity, loudness, direction*) had value. However, questions arose over which sound identities to filter and whether to trust the device's filtering decisions.

3.3.4 Social Contexts of Use

Regarding social acceptability, all but one participant (P6) felt comfortable using the smartwatch around other people, although some additional considerations arose. Many participants commented about not caring what others thought ($N = 9$) and/or were excited to show the watch off ($N = 6$). For example, P1 said, *"Any tool to help me access my surroundings is better than no tool, and, honestly, if people think it's weird, that's their issue,"* while P2 said, *"it wouldn't make you stand out, because most people are accustomed to some people wearing watches."* The watch was also seen as useful for spoken communication ($N = 11$); for example, *"If someone's behind me in a store, and they say, 'Excuse me,'"* (P14) and *"being able to pick up where a voice first starts coming in from"* (P2).

That said, eight participants discussed how their social context may impact usage. P3 said, for example, *"I'd be more likely to use it when I'm alone, because when I'm with my friends or my family, then I would depend on them."* Use in a Deaf cultural context was also discussed, with P16 acknowledging that sound awareness technology may not be appropriate around other Deaf people, a finding that reflects past work [37]. Similarly, P5 said: *"No matter who I'm with, I'd like to have the environmental information. [... But] with hearing people, they're trying to speak with you. Whereas around deaf people, they're signing."* Finally, six participants mentioned not wanting to negatively impact others by appearing distracted by the watch or vibration notifications.

3.3.5 Privacy Concerns

We asked participants if they had any privacy concerns using a watch with an always-on microphone—none did. P15 argued that the smartwatch supports user privacy because of its unobtrusive and commonplace

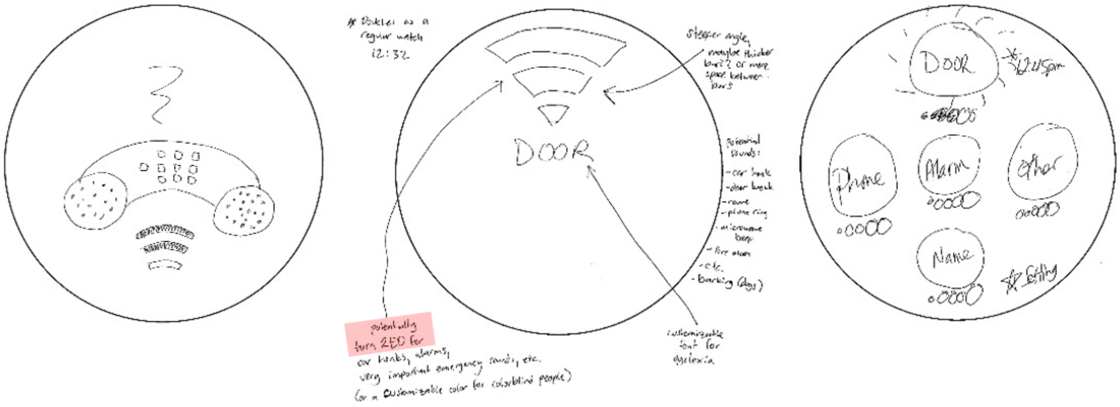

* Visible as a regular watch 12:32
DOOR
steeper angle, maybe thicker bars? or more space between bars
potential sounds:
- car honk
- door knock
- name
- phone ring
- microwave beep
- fire alarm
- etc.
- barking (dog)
potentially turn RED for car honks, alarms, very important emergency sounds, etc. (or a customizable color for colorblind people)
customizable font for dyslexia
Door
12:45pm
Phone
Alarm
Other
Name
Setting

Figure 3.7: Three example participant design sketches: icons for sound identities, such as a telephone (left, P5), use of colors and other changes (center with color words emphasized, P1), and co-occurring sounds (right, P6).

interaction: *"Because all you have to do is turn your wrist."* We did not probe for data collection and security issues, and the topic emerged for only two participants, such as: *"Where does all that information go? I would be interested in knowing"* (P12).

3.3.6 Design Suggestions

Participants provided design suggestions throughout the study, including user interface ideas, alternate haptic methods, and customization support. Though participants were not asked to critique our visual design, responses were generally positive—*e.g.*, P14 found it, *"very easy to read"* and *"very clean."* For new user interface ideas, participants suggested icons, having the screen flash or change brightness to indicate some sound characteristic, and the history of detected sounds. Adding color ($N = 7$) was most commonly requested to improve glanceability, encode other sound information, or *"for privacy reasons"* (P6). For example, P10 compared color to words: *"It's faster. [...] Like if it's blue, it's like, 'This is happening.' If it's red, 'This may be important.'"* Participants were also invited to sketch out design ideas; five chose to do so. Figure 3.7 shows three example ideas for iconography, color, and visualizing co-occurring sounds. There were fewer haptic-related ideas; the most common suggestion was to adjust vibration intensity to convey sound information ($N = 5$). Other ideas included using Morse code for tactons, providing directional information through multiple vibration motors, and continuously vibrating for emergencies.

For customizability, participants wanted to prioritize specific sounds, switch between filtering options, create personal tacton sets, and add preset settings for specific contexts. Responses were mixed when participants were asked if the watch should automatically adapt to different contexts. Several participants ($N = 7$) were receptive to the idea, suggesting that the watch could adapt based on location, noise level, or the wearer's activity. P8 said, for example, *"If it knows I'm driving, that would be great."* Some participants who were against the idea were worried about being confused by the changes ($N = 3$) and/or cited negative experiences with automatic adjustment in other devices ($N = 4$). P5, for example, mentioned that her hearing aid had *"...just kept changing on its own and filtering out sounds."* To mitigate these issues, several participants suggested the ability to override any automatic changes ($N = 4$).

Summary. The smartwatch form factor generally met user expectations regarding privacy and social acceptability; however, usage may vary depending on the social context (confirming [37]). Users preferred glanceable visuals, suggesting icons or color-coding in the design. Customization was considered important, but opinions were divided on whether settings should adjust automatically or require manual input.

3.4 Discussion

This study confirms DHH users' preferences for having both visual and haptic feedback in a wearable sound awareness system [37, 102, 104], but also: (1) extends our understanding of how to design these feedback modalities in combination, (2) demonstrates the potential for small sets of haptic patterns, and both (3) highlight user reactions to soundscape complexity in busy environments as well as (4) identifies promising methods for filtering that complexity (*i.e.*, based on sound characteristics and context). Here, we reflect on combining visual and haptic feedback for smartwatch sound awareness feedback, considerations for managing soundscape complexity, and limitations of our work.

3.4.1 Complementary Roles of Visual and Haptic Feedback

Visual and haptic feedback offer complementary roles for wearable sound awareness systems, and their combination provides users with flexibility. Advantages of visual feedback include high information throughput and ease of understanding. However, the small smartwatch screen is limiting, so simple and glanceable designs are preferred. Suggestions for increasing glanceability include using icons or color to encode sound identity—changes that could also be designed to preserve the wearer's privacy in the presence of others. Further, the visual designs in our study showed only a single sound at a time, with no notion of history. How (and if) to provide this more complex information on the watch is an open question.

Our study shows that haptic feedback is critical, as DHH people use visual cues for environmental awareness [98]. Haptic notifications (whether simple or pattern-based) are thus important because they attract the user's attention without interfering with existing visual awareness strategies. A related benefit is that haptic feedback could be particularly useful for safety-related notifications while the wearer sleeps. Despite these positives, however, overly frequent or obtrusive vibrations were seen as problematic, reflecting early work on tactile sound awareness systems [111]. In noisy situations, it may be best to allow users to turn off or reduce the haptic feedback and provide more visual descriptors of the soundscape.

Many projects [53, 112] have explored tactons for haptic communication, though our work is the first to apply them to sound awareness. While participants expressed concern about the learning curve and the time required to interpret a tacton, the overall response was positive. Due to our limited control over our smartwatch's vibration output, our preliminary designs were meant to assess the general idea of using tactons. Thus, future work must focus on more specific design attributes, such as intuitive tacton sets. Our findings suggest that tactons should be limited to a small set and that, due to individual preferences for what information to convey via tactons (*identity*, *loudness*, *direction*), users should have the ability to configure how they are used.

3.4.2 How to Manage Soundscape Complexity

The sheer complexity of the *in situ* soundscapes impacted participant responses to sound awareness feedback. While a previous survey showed that only 63% of DHH respondents predicted they would want sounds filtered [37], all participants in our study desired filtering—a disparity we attribute to our participants' exposure to realistically complex soundscapes. To manage this complexity, past work has limited notifications to specific sounds [97] or, for vibrotactile feedback alone, to sounds above a loudness threshold (*e.g.*, [133]). Our study is the first to compare different filtering options and to examine filtering based on direction. Positive responses to all three options (*identity*, *loudness*, *direction*) and varied ideas about how each would be useful suggest that future work should continue to evaluate these options and to further refine their designs. Specifically, future work should examine more realistic sound pacing than our study's consistent stimuli pace (one sound every 3 seconds) and explore the ability to switch between filtering presets depending on the context.

The feasibility of the filtering designs we evaluated is an important factor to consider. Filtering based on *loudness* can be done with a simple threshold amplitude level, though our findings suggest that the threshold may need to automatically adapt to background noise levels and/or be controllable by the user. In contrast, sensing sound *direction* is more difficult and would likely need additional hardware, such as a wearable microphone array [76]. Reliable *identification* of open-ended sounds is also complicated by overlapping sounds, background noise, and differences across locations [13, 44]. For sound identification, our study focused on reliably identifying a small set of sounds, as demonstrated by Bragg *et al.* [13].

Automatic filtering also introduces ethical and practical considerations of how much trust a DHH user should put in a sound awareness system and what constitutes an appropriate and accurate representation of the surrounding soundscape. Trust, for example, was highlighted by P4 in our study, who preferred to rely on her existing sound awareness strategies than trust a system in unfamiliar locations. An important complication is that a DHH user may have limited means of judging the sound awareness system's accuracy for themselves, such as noticing errors in identification or filtering. While researchers will need to continue grappling with these issues, system transparency offers a potential path forward: systems should be transparent in making decisions, provide real-time information about what is being filtered/identified, and allow users to modify those decisions as necessary.

3.4.3 Limitations

We enumerate three primary limitations. First, our volunteer participants may have been biased toward sound awareness technologies compared to others in the highly diverse DHH population; for example, larger surveys of DHH participants show that some segments of the population are less interested than others [13, 37]. Second, limited exposure to tactons may have reduced their perceived utility compared to if participants had had more time to learn and use them. Third, our *in situ* exploration was brief, did not show real sounds, and occurred within a small radius. While this setup allowed us to identify new considerations that purely lab-based evaluations had not previously seen, work in other contexts and study of longer-term use is needed to understand the tool's broader utility and adoption/abandonment issues.

3.5 Chapter Summary

Context and physical location have a strong influence on the sound awareness needs of DHH users. Smartwatches are the preferred device for portable sound awareness among DHH users, but prior work has yet to examine effective sound feedback on these devices. In this chapter, we used the design space smartwatch-based sound feedback to examine DHH users' preferences for receiving sound information and the influence of contextual factors on these preferences. A Wizard-of-Oz study with 16 DHH participants revealed a strong preference for sound information through combined visual and haptic (vibration) feedback, with benefits and tradeoffs between simple vibrations or informational vibration patterns (tactons). Participants emphasized the importance of identity over other sound characteristics (direction, loudness), and all participants desired filtering options to manage sound feedback in real-world soundscapes. A strong majority wanted to filter by the sound's identity, but they did not agree on which identities to filter and whether to trust the device's filtering decisions—highlighting a need for tools that are adaptable to changing contexts and individual preferences. This chapter informs the design of sound recognition tools' *output*; the next chapter shifts focus to their *input*, exploring how DHH users capture and interpret data from real-world soundscapes to build personalized tools.

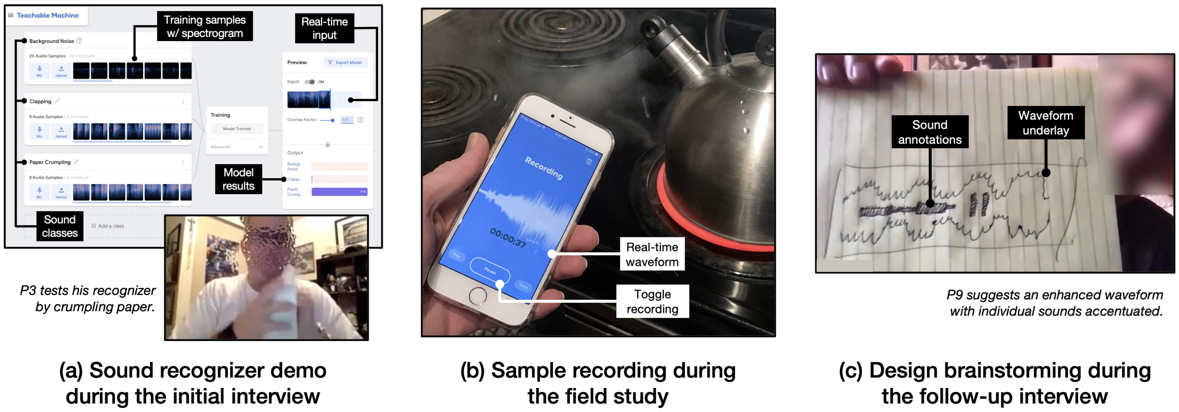
Chapter 4

Investigating Real-World Audio Data Collection[1]

4.1 Introduction

The last chapter's exploration of smartwatch-based sound feedback highlighted how sounds' identity is valuable information for DHH users across a variety of contexts. Work in sound awareness [13, 46, 71, 97, 98] shows that DHH users desire sound recognition to augment personal safety (*e.g.*, footsteps) and social awareness (*e.g.*, nearby voices), and to respond to non-urgent alerts (*e.g.*, home appliances). To meet these needs, automatic sound recognition features are now included on both major mobile platforms: Apple iOS [7] can notify users when it recognizes eleven sound categories (*e.g.*, baby crying, car horn), while Android's Sound Notifications feature [51] supports ten sounds plus a timeline of all recently detected sounds. However, these features—and prior work implementing sound classification for DHH users [71, 72, 103, 127]—use generic models that are pre-trained on large sound corpora for a rigid set of sound classes, and as a result may not adapt to user-specific needs.

[1] This chapter includes materials originally published in [47], which explored the experience of DHH users when recording sound samples to personalize a sound recognizer.



(a) Sound recognizer demo during the initial interview

(b) Sample recording during the field study

(c) Design brainstorming during the follow-up interview

Figure 4.1: We conducted a three-part study with 14 DHH participants. In Part 1 (a), we conducted an initial interview and participants recorded samples of *clapping* and *paper crumpling* to train Teachable Machine [17, 49], an online sound recognizer. In Part 2 (b), participants recorded sounds in the field for one week. In Part 3 (c), we conducted a follow-up interview and brainstormed design ideas for specialized feedback.

Designed for universal support, this "one-size-fits-all" approach to sound recognition does not meet DHH users' requests for personalized sound categories (*e.g.*, family members' name calls [13, 71]) nor does it account for edge cases in real-world sound events (*e.g.*, a generic cat *vs.* my cat). A potential solution is to incorporate approaches from human-centered ML research [36, 118] to support DHH users in training personalized models of their own. However, end-user ML solutions that augment human sensory abilities present a unique challenge for users who have sensory disabilities [38, 75]: how can a DHH user who has difficulty hearing a sound themselves effectively record samples to train an ML system to recognize that sound?

Building on work by Kacorri *et al.* and others (*e.g.*, [75, 88, 129]) to support blind and low-vision people in training personal object recognizers, we explore the parallel question of how DHH people can train personal sound recognizers. In contrast to the rich corpus of blind photography work (*e.g.*, [1, 73, 141]) that underpins the visual object recognizer efforts, very few studies have focused on how DHH users record and engage with audio data—despite this data predicating a sound recognizer's effectiveness for DHH users. One exception comes from Bragg *et al.* [13], who surveyed DHH people on their sound awareness needs, used the findings to design a personalizable sound recognition prototype, then ran a brief Wizard-of-Oz study where DHH participants recorded samples of two sounds (alarm clock, door knock) to train a model. Another exception is a workshop study by Nakao *et al.* [107] that had DHH participants collaboratively interact with a sound recognition interface to characterize their understanding of ML, such as challenges with defining ML tasks for sounds they know but cannot hear. Both studies demonstrated the potential for DHH users to train a sound recognizer; however, several open questions remain; for example, what considerations do DHH users make when recording in environments with real-world acoustic variation—like overlapping sounds and background noise [91]—and what kinds of features can aid DHH users in assessing their recorded samples as training data?

To understand the experience and needs of DHH users in recording sound samples to train future personalized sound recognition systems, we conducted a three-part study with 14 DHH participants:

1. An initial interview session to provide an introduction and hands-on engagement with an existing personalizable sound recognizer [49];
2. A week-long field study to independently record sounds of interest via a smartphone app;
3. A follow-up interview to discuss the experience and design probes for new recording and training tools.

We focus our analysis on considerations made while approaching the recording task, perceived challenges and successes during recording, and interpretations of the quality of recorded samples. Participants conveyed a positive outlook towards these tasks and felt most confident recording sounds that were continuous, prominent, and controllable (*e.g.*, a faucet). However, they described challenges in recording spontaneous, invisible, or complex-to-produce sounds (*e.g.*, emergency sirens) that could make training important sound categories infeasible for DHH end-users. Participants often considered their data in terms of its diversity—reflecting prior work with other non-expert ML users [107, 148]—but their limited auditory experience led to unique challenges in determining the diversity among their samples, as well as how representative each sample was to its real-world counterpart. These and other challenges resulted in several design suggestions for more specialized feedback.

This chapter contributes: (1) an empirical account of non-expert DHH users' experience with real-world audio recording to train a personal sound recognizer; (2) characterization of DHH users' conception of

Table 4.1: Demographics of study participants. HH = hard of hearing.

ID	Age	Gen.	Iden.	Hearing loss	Hearing dev.	Relationship to sound	ML exp.
P1	20	W	HH	Profound	Both	*"I assume the same way that hearing people perceive sound (when I use my cochlear implant and hearing aid), but with more mental concentration. As well as some gaps, like not noticing or picking up quieter and/or unclear sounds."*	Some
P2	53	W	Deaf	Profound	Hearing aids	*"With an assistance of hearing aid, I learn to identify the sound based on the vibration an/or the rhythm. I hear the pitch, note, timbre, range... but I can't identify the spoken words."*	Slight
P3	47	M	Deaf	Profound	Hearing aids	*"When I hear sound by my hearing aid, I can feel that sands jump in my head."*	Some
P4	23	M	HH	Mod. Severe	Hearing aids	*"With hearing aids on, I experience sound much as a hearing person would, with maybe a bit more difficulty. Without hearing aids, sounds are kind of muddled and muffled, leaving me to parse together words based on mouth movements, context, and location."*	Slight
P5	56	W	Deaf	Profound	Hearing aids	*"Environment sounds help me know what is going on."*	Slight
P6	24	W	Deaf	Profound	Cochlear imp.	*"I wear my two cochlear implants to listen to the sounds. [...] I can hear music [with] words that I don't understand, and I can understand the sounds around me, such as alarms, television, conversations from people."*	Some
P7	28	M	deaf	Profound	None	*"I was born to live without sound, so I never really knew what the sound is all about. Music is probably the loudest thing that I can relate to, even though, I can't hear it at all, just the vibrations."*	Some
P8	87	M	deaf	Profound	Hearing aids	*"I experience voice through my hearing aids directly or through my mobile phone. Other sounds in the world are muted or absent."*	None
P9	69	M	Deaf	Severe	Hearing aids	*"I use it for language, as English is my first language. I don't listen to music. I prefer to not use my hearing aid at home, unless I'm watching TV."*	Slight
P10	70	W	HH	Mod. Severe	Hearing aids	*"[Sound] is always distorted and I don't know which direction it is coming from."*	None
P11	44	W	Deaf	Profound	None	*"I rely on vibrations [...] [and] visual alerts (looking outside my window for expected deliveries or someone arriving at my destination), and mostly have few people informing me of the sounds."*	Some
P12	35	W	Deaf	Profound	Hearing aids	*"I'm full Deaf so most sounds don't make sense to me."*	None
P13	19	M	Deaf	Severe	Hearing aids	*"I can hear sound very quietly without my hearing aids and with it it becomes amplified but I can't process the sound correctly."*	Slight
P14	31	W	Deaf	Profound	None	*"I need a tool that acknowledges important sounds or noises."*	Some

real-world audio data in an ML context, including sense-making strategies; and (3) design implications to support DHH users in building their own personalized sound recognition systems.

4.2 Methods

To understand user experience when recording sound samples for a personalizable sound recognition system, we conducted a three-part study with 14 DHH participants: an initial interview session, a week-long field study to record samples, and a follow-up interview and design probe activity.

4.2.1 Participants

We recruited 14 DHH participants via email lists at two U.S. universities and via social media and snowball sampling (Table 4.1). Eight participants identified as women, and six identified as men. Participants were, on average, 43.3 years old (*SD*=21.3, *range*=19-87). Nine participants identified as Deaf, three as hard of hearing, and two as deaf. Ten participants reported using hearing aids; two used cochlear implants; one used both. We required access to a laptop or desktop computer, a stable internet connection for video conferencing, and a smartphone with 150MB in free storage for recording sounds during the field study. Informed by Hong *et al.* [59], we asked participants to rate their familiarity with ML on a four-point scale: three reported never having heard of it (*not familiar*), five had heard of it but did not know what it does (*slightly familiar*), and six reported being *somewhat familiar* with what it is and what it does. No participant reported having extensive knowledge of ML (*extremely familiar*)—indicating our participants were non-experts. After initial interviews with six participants, we added two technology-related screening requirements: using a laptop or desktop computer at least once a week and using a smartphone for tasks other than phone calls and text messaging at least multiple times a week. Participants received a $125 gift certificate as compensation.

4.2.2 Procedure

The study had three parts: an initial interview session to introduce audio recording for sound classification, a one-week use of an audio recording application, and a final interview and design probe session (Figure 4.1). Participants also completed an online pre-study questionnaire to collect demographics and gather information on sound support technologies, general technology familiarity, and their perspective on important sounds in daily life. Consent forms were emailed to participants in advance and verbal consent was taken at the start of the initial interview session.

All interviews were led by the first author and held remotely using videoconferencing software. Participants could request their choice of accommodation: nine opted for sign language interpretation, and, two opted for real-time captioning, three opted for no accommodation. Before the study, we shared the interview materials in an online slide deck (see Supplementary Materials) and employed the videoconferencing's "Share screen" feature. During both sessions, connection problems caused P7's ASL interpreter to drop out for several minutes; we continued the discussion via the videoconferencing chat feature.

Participants received non-auditory feedback via waveform and spectrogram sound visualizations during the initial session (Figure 4.2) and the waveform alone during the field study (Figure 4.1b). Waveforms show the amplitude—or loudness—of sound over time and are common in audio recording, editing, and playback software. DHH participants in prior work liked waveforms while recording samples in a lab setting [13]; we explore their value for samples recorded in daily life. Spectrograms show the frequency spectrum over time, are often used for scientific analyses (*e.g.*, bioacoustics [25]), and can be difficult to interpret for novice hearing users [18, 63]. Early work showed frequency information was inadequate for DHH users in a sound identification task [98]; we briefly explore DHH participants' opinions of spectrograms for displaying sound activity. Below, we detail the three sessions of the study.

Initial Session (75 min)

The initial session began with 15 minutes for setup and orientation, followed by a discussion and demonstration of how to personalize a sound classification tool. We provided a definition of ML in an audio context, described the possible benefits of a trained model using personal recordings, and asked participants about their prior experience with audio recording.

Then, to provide hands-on experience with a personalizable sound recognition tool, we introduced Google's Teachable Machine for audio [17, 49] and its spectrogram visualization (Figure 4.1a). We led the same discussion with all participants during this activity, but only ten of the 14 successfully trained the model themselves; the other four experienced technical difficulties and instead watched the recording and training process on the study coordinator's screen. Participants trained three sound classes: *background noise* as required by Teachable Machine (*i.e.*, *"typical sound activity"* in the current setting), *hand claps*, and *paper crumpling*. We chose these two classes because they are produced by simple physical actions, are reproducible (to provide multiple samples to the machine), and have visually distinct frequency signatures in their spectrogram representations. We used the videoconferencing annotation feature to explain Teachable Machine's interface but allowed participants to record samples independently and delete and re-record for any reason. We instructed participants to produce each sound continuously for several seconds (*e.g.*, *"clap your hands"*), then use Teachable Machine's extraction feature to split the recording into one-second samples—the required sample format.

After collecting the minimum samples required by Teachable Machine for each class—20 for *background noise* and eight each for *hand claps* and *paper crumpling*—we invited participants to share their interpretation of the spectrogram audio representations.[2] The data was then passed to Teachable Machine's training module to construct a working classification model. To demonstrate both the capabilities and limitations of the tool, we instructed them to test the model by again clapping their hands and crumpling paper and to produce other sounds that the tool had not been trained to recognize (*e.g.*, knocking on the table).

Following the Teachable Machine demonstration, we discussed possible characteristics of high-quality sound samples for training a sound recognizer. Informed by training datasets used in prior work [71, 87], we provided a list of five desirable characteristics to guide participants during the field study (see Supplementary Materials for complete instructions):

- **One sound per sample**: The targeted sound is present and louder than other sounds in the sample.
- **Appropriate background noise**: Other noise in the sample should be typical of noise in that location.
- **Accurate labeling**: The sample is named after the contained sound.
- **Personal**: The sample replicates how the sound occurs in your daily life.
- **Complete**: The sample contains the entire sound from start to end.

While pre-processing algorithms may be used to separate multiple sound sources [94] or remove background noise [26], we included both in our guide to prompt consideration of auditory phenomena that may otherwise not be apparent to DHH people.

To further spur participants to consider how sounds are captured in a recording, we presented five video clips of realistic sound scenarios (Figure 4.2, from left): *"tea kettle whistle in a quiet home"*, *"baby crying during a thunderstorm"*, *"emergency siren passing on a busy street"*, *"dog barking outdoors on a summer day"*, and *"door knock during a small party"*. After each video, participants provided their own interpretation of each sound scenario's waveform and spectrogram, and then the hearing first author connected salient areas

[2] We hoped to prompt consideration of how sound activity can manifest visually. Responses were not included in our analysis.

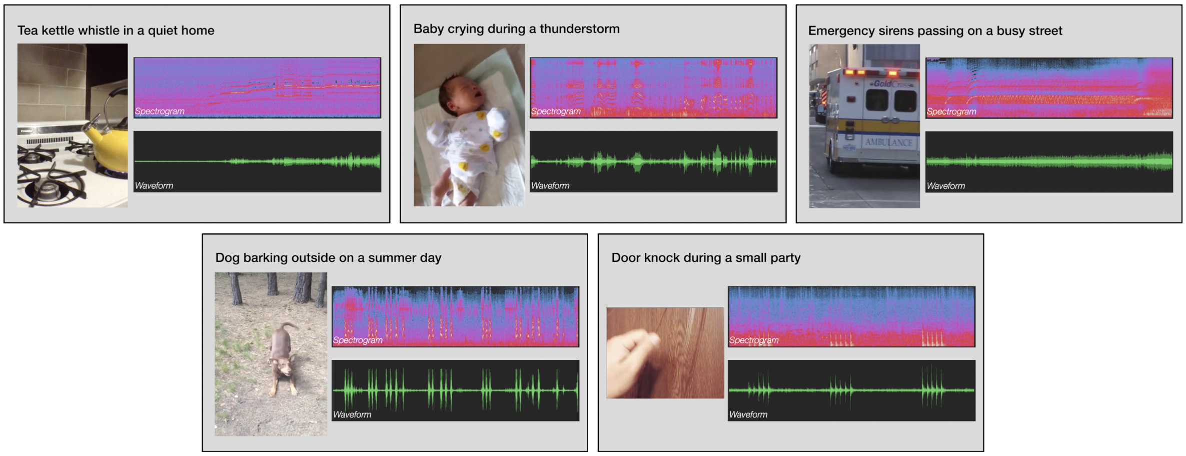

Tea kettle whistle in a quiet home
Spectrogram
Waveform
Baby crying during a thunderstorm
Spectrogram
Waveform
Emergency sirens passing on a busy street
Spectrogram
Waveform
Dog barking outside on a summer day
Spectrogram
Waveform
Door knock during a small party
Spectrogram
Waveform

Figure 4.2: Videos shown to participants to introduce real-life recording challenges and visualizations. Spectrogram and waveform visualizations were generated using Audacity [8] and set to advance in sync with the video clip. To match the described context of the *baby crying*, *dog barking*, and *door knock* events, the hearing first author selected an additional audio file (*e.g.*, a recording of a thunderstorm) to layer on top of video's audio.

of each visualization to events in the video (*e.g.*, thunderclaps during *baby crying*). We used a comparison slide with all five sound scenarios at the end to solicit participants' overall opinions of the spectrogram and waveform visualizations. The session concluded with instructions and setup for the field study, as described next.

Field Study of Recording Practice (1 week)

To study how people who are deaf or hard of hearing may record audio samples to train a personalized sound recognition system, we asked participants to record sounds in their daily lives for a week. At the end of the initial session, we helped participants download and configure the *Rev Recorder* app [119] on their smartphones. In preparation for our study, we reviewed various smartphone-based sound recording apps and selected Rev because it has a simple, well-designed interface with high-contrast waveforms, provides immediate cloud backup of recorded clips, and is free on Android and iOS. We set up Rev to automatically upload recordings to an anonymous Dropbox account created for each participant. We also explained how to temporarily disable this "auto-upload" function when the recording might capture sensitive information.

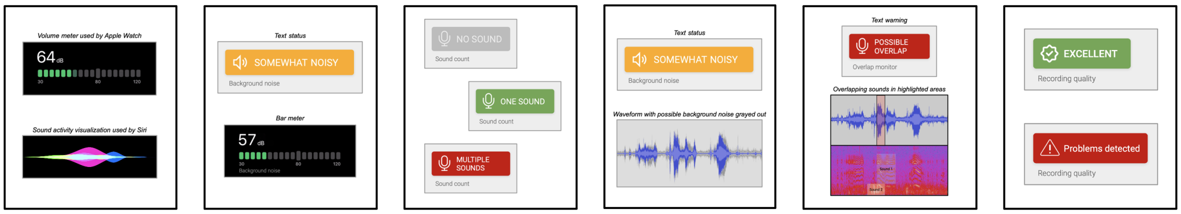
For the recordings themselves, we asked participants to record at least three different non-speech sounds each day for at least five days over the week (*i.e.*, at least 15 unique sounds in total). To respect participants' time, we imagined the training set would follow a few-shot learning approach: we asked for three to five samples of each sound if possible, with exceptions allowed for sounds that may not occur often (*e.g.*, an ambulance siren). We recommended that samples be 5-10 seconds long, but we did not provide a strict time limit for flexibility. We allowed participants to record samples of *any* non-speech sound, though we asked them to prioritize recording sounds that they thought would provide value in a sound recognition tool. While some DHH users may be able to ask hearing people for support, others may not, and we requested that participants not ask other people to help with the recording or to share input on the quality of a sample to learn about independent recording experience as a baseline. Participants could, however, ask another person to produce the sound needed for a recording (*e.g.*, asking a friend to knock on the door).

Each day, participants were prompted via email or text message at a pre-arranged time to complete an online diary questionnaire. This questionnaire asked for: a list of the sounds recorded that day; motivation for recording those sounds; successes and challenges during recording; attempted but unsuccessful recordings; and additional information that would have helped with the day's recording.

Follow-up Interview and Design Probe Activity (60 min)

We scheduled a final video call during which the interviewer screen-shared a new slide deck. We provided a copy of the participant's diary entries to reference as needed throughout the session. The first half of this session consisted of a semi-structured interview on the participant's overall experience with recording sounds, any contrast between that and their initial expectations, and whether they had changed their recording practices over the course of the week. Next, we provided a complete list of the sounds they had recorded for review, and we asked them to identify the easiest and hardest sounds to record and if they were satisfied with their recordings. Finally, we reviewed the list of high-quality recording characteristics and discussed how each factor surfaced (if at all) during the week.

The second half of the session consisted of a design probe activity inspired by Hutchinson *et al.* [62] to discuss new ideas for supporting DHH people in independently sampling sounds. We first asked the participants to describe their ideal features, then presented ten possible feature ideas (Figure 4.3). For each idea, we showed a brief description and two mockups. We asked whether each feature would be useful and



(a) Volume monitor (b) Background noise monitor (c) Overlapping sound monitor (d) Background noise feedback (e) Overlapping sound feedback (f) Quality rating

Figure 4.3: Six of ten sampling features taken from the example slides shown to participants. Not shown: waveform, spectrogram, background noise removal, and trimming function. (a) Volume monitor, (b) background noise monitor, and (c) overlapping sound monitor were examined for real-time support; (d) background noise feedback, (e) overlapping sound feedback, and (f) quality rating were for post hoc sample review.

if the participant had any related design ideas. Finally, we displayed a list of the ten features and asked which to include in a redesigned recording app, if any essential features were missing, and for the single most essential one. We concluded the session by asking how this app might have changed their experience recording samples during the previous week.

4.2.3 Analysis and Positionality

Using reflexive thematic analysis [14, 15], we iteratively coded transcripts of both interview sessions and responses to the field study reflection form. Our analysis was semantic and realist, and we developed themes using a mixed inductive and deductive approach; for example, we structured broader theme development around the steps required to personalize a sound recognizer, but we organically identified themes within each step. The first author briefly read through the data, generated initial codes, and then applied these codes to data from two randomly selected participants. Another researcher reviewed the code applications and then met with the first author to further discuss and refine the codes. The first author coded the remaining transcripts and generated themes from data excerpts collated from each code. A reflexive approach to thematic analysis emphasizes findings actively shaped by the research team's social, cultural, and academic biases. The first author—who ran all interviews and led analysis—is hearing. Some authors—involved in study design, analysis, and writing—are Deaf or hard of hearing. All research team members have backgrounds in human-computer interaction, and many are computer scientists by training.

4.3 Findings

We begin with a quantitative overview of participants' recordings from the field study, followed by a report on their ML expertise. Then, we synthesize their experience based on two key ML components requiring subject matter expertise (informed by Yang *et al.*'s study of non-expert ML users [148]): (1) data and label collection, examined through participants' overall approach to recording training samples; and (2) data interpretation, examined through their assessment of samples' contents.

Occasionally, we include quotes that demonstrate confusion on the part of a participant, perhaps due to a misconception of sound or a misunderstanding of the feedback visualizations themselves (*e.g.*, a participant suggests that they could determine pitch from a waveform visualization, which is not possible). We mark these quotes where relevant.

4.3.1 Overview of Recorded Samples

The 14 participants recorded 677 sound samples in total during the one-week field study ($M = 48.4$ per participant, *SD*=23.3, range=13-86). They provided 243 sound classes (Figure 4.4) at an average of 17.4 classes per participant (*SD*=5.1, range=10-29) and 2.8 samples per class (*SD*=1.2, range=1-10). We used the `pydub` library [120] to analyze each sample's duration, average loudness in decibels relative to full scale (dBFS), and silence—defined as any period of 1s or longer where the amplitude was 16 dBFS below the file's average. Samples averaged 11.5s in duration (*SD*=4.6, range=2.4-34.8). The average loudness of each sample was -34.1 dBFS (*SD*=9.5), with P9's *"Tea kettle whistle"* (3 samples, -15.9 dBFS) being the loudest class by average and P2's *"Bathroom"* being the quietest (1 sample, -64.7 dBFS). Regarding silence, 158 samples (23.3%) contained at least one silent period lasting 1s or more, and 78 of these (11.5% of the total set) contained 3s or more of long silence(s). The length of long silence in each sample was, on average, 3.6s (*SD*=2.3) and 36.6% of the sample's total duration (*SD*=22.3%).

To compare the contents of each sample with its label, we randomly selected half of each participant's samples (*N*=338) and rated *yes* or *no* if the labeled sound class was heard in playback, or *unclear* for ambiguous or unfamiliar sounds. This analysis was meant as a brief, subjective inspection from a hearing user's perspective and not a full assessment of the samples' overall quality for training an automatic sound

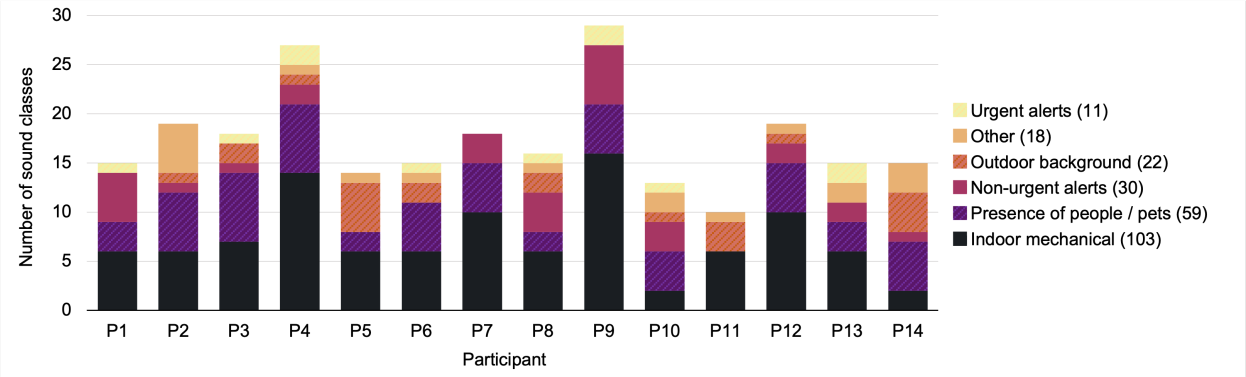


Figure 4.4: Breakdown of sound categories recorded by participants. The total classes in each category are shown on the right.

recognizer. Two hearing researchers independently rated 52 of the samples, met to resolve disagreement and formalize a rating scheme, and one of the researchers completed the remaining set. The labeled sound was heard in 92.0% of the samples and missing from 3.6%; the remaining 4.4% were rated *unclear*. For example, P4's *"Freeway traffic noise"* and P5's *"Car on street"* were rated *yes* for prominent vehicle sounds, P14's *"Busy street noise"* was rated *no* for near silence, and P12's *"Car running"* was rated *unclear* for an ambiguous droning sound. Other *unclear* sounds included P2's *"Friend's apartment"* and P14's *"Oil and garlic"*.

4.3.2 ML Expertise and Prior Recording Experience

Participants demonstrated a range of ML knowledge in the first session despite all being non-experts. P4, for example, recalled a lesson on ML in *"one of [my] data science classes"*, while P8 was a newcomer: *"This is brand new to me, but I kind of get the idea."* P7 described practical ML applications, including media recommendations (*"Netflix uses it"*) and automatic speech recognition to communicate with hearing people: *"Sometimes you'll need to have a human interpreter [...] but it can be nice to have the speech recognition that you could use in an emergency."*

Following our brief explanation, all participants showed an approximate comprehension of user-driven sound recognizer personalization. For example, P3 described the machine's workflow as, *"I make a sound, or there's a sound [happening], and this device will copy it? And then later, when the sound is repeated, it will tell me what it was?"* Participants also recognized the risk of misclassification errors, although P12 was interested in having agency in fixing them: *"Machines aren't perfect, and they can make mistakes too, [...] but I don't mind that. I'd like to help the phone app [to learn]."*

Most participants had recorded audio before, such as for song identification (*e.g.*, Shazam) (P6, P7, P13), to play for hearing people (P2, P3, P12), and to capture school lectures (P2, P14). However, this experience was limited. For example, P8 had only briefly recorded himself playing guitar, while P4 said his experience was incidental: *"It's when I'm recording video, and there happens to be audio [with it]."* P3 recalled learning from a hearing person on a video call that a smoke detector in his house was beeping to indicate low battery, so he recorded all of his smoke detectors and shared the recordings to figure out which battery to replace. P12 had also sent recordings to hearing people, *"Just to make sure something's not left on,"* while P2 had recorded sounds for fun to *"test"* her hearing friends' sound recognition abilities.

Despite limited recording experience, all participants were enthusiastic about recording to train a personal sound recognizer during the initial session. Every participant shared at least one desired sound for an automatic sound recognizer; interests primarily focused on urgent and social sound, in line with prior work [13, 37, 99]. Examples included fire alarms (P4, P5, P9, P14), leaking or running water (P7, P9, P10, P12), musical instruments (P2), and name calls from a partner (P9). P10 wanted to know if she had left her car running: *"[If] there was a warning on your iPhone because it was still hearing the sound of the engine [...] that would be awesome."* Participants also shared ideas about what might make recorded audio samples better or worse for training their sound recognizer; better samples were assumed to have *"clarity"* (P5) and be *"loud enough"* (P8) while worse samples could be affected by *"overlapping sound"* (P6) or *"white noise"* (P1, P5).

4.3.3 Planning and Recording Samples

We now describe what sounds our participants recorded in the field, how they planned and executed these recordings, and the approaches they used to interpret these samples.

Selecting Sound Classes to Record

Participants primarily chose sound classes that were personally meaningful or a source of curiosity, although some sounds were recorded out of convenience. Meaningful sounds generally aligned with those identified during the first session and in prior work [13, 37, 99]—urgent alerts, social presence, and home appliances. For example, P14 recorded sounds from her pets to warn *"if something had happened to them,"* while P2 chose to record doors *"to know [if] someone is in the apartment."* Curiosity toward a sound—such as *"a music box"* (P10) and *"ocean waves"* (P11)—motivated other choices, although these were likely due to the novelty of the recording activity rather than imagined use cases for a recognizer. Many choices emerged organically over the course of the study; as P5 described it, *"The 'doing' became more interesting as a result [of recording]—the more I did, the more I wanted to do."*

Our study instructions and physical constraints from the COVID-19 pandemic limited some sound choices.[3] We prohibited recording speech for privacy reasons, although P9 disregarded this instruction to record his partner calling out his name *"in an emergency"*. P14 recorded an online video of an emergency siren rather than the real-life source (contrary to our request), explaining: *"I couldn't stand outside and wait for one to come by."* Although no participant mentioned pandemic-related social distancing guidelines as a serious limitation, most chose to record all sounds in and around their home.

Considering Decision Boundaries and Diversity

When defining each sound class, participants reported considering possible decision boundaries and the appropriate diversity across samples, but they described uncertainty due to their limited perception of each sound. With regard to decision boundaries, P2 wondered if *"a kitchen fan and the bathroom fan"* sounded different enough to allow separate classes, while P9 imagined that a faucet in *"a stainless steel rectangular sink"* and *"a rounded porcelain sink"* might sound different enough to allow for separate classes to convey each faucet's location. P9 further estimated that the faucet *"running"* and *"dripping"* would necessitate separate labels despite being uninterested in that distinction himself: *"I just want to know [the faucet] should be turned off."* Other participants hoped the machine could inform them of nuanced sound information, but

[3] The study was carried out during the summer of 2021.

they did not know how to convey this nuance through their data; for example, P1 only recorded one *door closing* class despite wanting more detail:

> *"Someone could slam it, it could be more aggressive, it could be like a soft one. [...] Seeing a sound recognition [tool] be like, 'The door closed.' I don't know if that's super helpful to me because it doesn't give me the nuance of information or what 'door closing' really is. [Maybe] someone's mad or maybe a window's open somewhere in the house that causes the door to slam shut."* (P1)

Likewise, P7 was enthusiastic about nuance in the sound of running bath water—*"It'd be nice to be away for a few minutes and come back when the sound is decreasing [...] to turn the water valve off"*—but only captured samples for a single *"bath water"* class.

Participants also considered the diversity of samples within each sound class—common among non-experts (*e.g.*, [58, 107, 148]). Many decided to limit diversity by producing the sound the same way in each sample: *"I want the sounds to be relatively consistent, just so the machine learning device isn't like, 'You have three different weird noises, but you say they're all the same'"* (P4). However, some attempted to vary the sound *"so the machine learning capability would be able to understand it more"* (P13). The hands-on experience with Google's Teachable Machine seemed to influence this thinking; for example, P2 wondered how the application would handle the real-life complexity of sounds: *"Some papers [are] heavy, some papers [are] light. [...] If you've already crumpled the paper and then try to re-crumple it, [then] that's going to be a different quality."* This motivated P2 to capture diverse samples during the field study; she wrote in one of her daily reflections, *"I suspect the doors and [blinds] sound differently when they are pulled or pushed in different speeds. It's good to have variation to help the recorder to recognize [doors] with different sound qualities."* However, this further emphasizes participants' uncertainty toward the real-world variation among the makeup of each sound class—highlighting an area where DHH users may need support.

Factors Impacting Sampling Difficulty

All participants successfully used Rev and described recording sounds as *"easy"* (*N*=9), *"interesting"* (7), and *"fun"* (P4, P10). Most described an initial learning curve that lessened with experience: *"Once I got used to it, I was able to record like a champ"* (P11). **Continuous** sounds were said to be particularly easy to record; for example, P12 said to record her floor fan, *"all you got to do is [...] just sit there with the app."* Other easy-to-record sounds were **prominent** (*e.g.*, *"microwave beep"*, P14) and directly **controllable** (*e.g.*, *"flushing the toilet"*, P13).

Uncontrollable sounds, such as pets' noises, required a different approach. For example, P14 struggled to anticipate her cat's activity: *"When would [it] purr? [And] predicting when it would meow [...] I had to kind of wait for them."* After failing to record his cat early in the week, P3 found a creative but unreliable way to elicit meows: *"I closed the office's door. [...] [My cat] was like, 'Meow, meow, meow! I need to get out.' [But] then the second time, she wouldn't meow. I had to let her out and then try it again."* P2 looked for visual signals to anticipate sounds from a friend's cat, like *"trying to wait until she opens her mouth"* to start recording.

Time-delayed sounds were easy for some to record because they followed a straightforward process: *"The tea kettle; I [only] had to wait a little while for it to boil—and the microwave signal; just turn it on for a few seconds [and] wait for it to stop"* (P8). Others found this inconvenient: *"I had to wait for the water to start bubbling before I could see it"* (P11).

Visual indicators were essential when recording spontaneous sounds, such as the arrival of *"the garbage truck"* to record *"the dumpsters [emptying] outside"* (P4). At times, this prevented sampling for otherwise desired sound classes: *"I couldn't record [a] bird chirping that was outside—I had no idea when to start the recording. And emergency vehicles—like sirens—if I wasn't able to see the vehicle, then I couldn't do it"* (P7).

Complexity in producing the sound was mentioned as another challenge: *"[I was] multitasking like, 'Did I turn it on? Is the app running? Is this going? Is the garage door okay? Am I going to get hit? What's going on?'"* (P12). A few participants recruited family and friends for help producing these sounds, but this introduced new challenges. For example, P1 avoided directing her father, as she worried it might suggest a lack of appreciation: *"I had to give it over to him, like, 'Oh you can create the sound. I don't want to critique you too much.'"*

Summary When defining their sound classes, participants considered possible decision boundaries and appropriate diversity for their samples, but inexperience with ML and the real-world variation in the sound population led to decisions based on guesswork. They described continuous, prominent, and controllable sounds as easiest to sample, but spontaneous, invisible, and complex-to-produce sounds were more difficult—even impossible—to capture.

4.3.4 Interpreting Sound Samples' Contents

During the field study, participants used Rev's waveform to visualize the contents of their samples. However, limitations with post hoc assessment strategies—such as audio playback and waveform comparison—caused participants to desire additional feedback.

Waveform Use

Participants liked waveforms' *"clear"* (P5) and *"not complicated"* (P2) design that could *"visually represent what is happening"* (P3) to *"see the rhythm"* (P5, P11). Several participants said it provided crucial support while recording samples; without it, P7 said he *"would have had no way of knowing that I was recording the sound right."* The waveform was commonly used for identifying concurrent or overlapping sounds by looking for *"some kind of 'off-pitch"'* (P13)[4] or anything *"unexpected in the shape"* (P1). One such unexpected noise came from P1's own physical activity, which she believed was unacceptable for a training set: *"Touching a doorknob; that touch kind of creates a sound. [...] It showed up very obviously in the waveform and I was like, 'Oh, I've got to re-record it."'* However, while P14 liked watching the waveform while recording, she could not use it to identify concurrent sounds: *"Some sounds were noisy certainly, but [...] [any] overlapping sounds were hard to distinguish and separate out."*

Despite the waveform's positives, the visualization did not always align with participants' intuition of sound and led to breakdowns in use. For example, P6 expected to see large peaks for thunder when recording a storm but found a *"jumble of noise"* and a *"blob of information"* that confused her.[5] To overcome this, she requested the waveform *"at least tell me what's higher and lower frequency."* At times, participants' residual hearing ability allowed them to mitigate waveform breakdowns; for example, after P14 *"put the phone right on the cat [...] and it didn't really look like it was purring"*, she concluded, *"Some of the things were too quiet and they weren't able to be captured."* However, after P1 noticed an empty waveform, her residual hearing ability allowed her to discover, *"If you replay, you can just make out the water dropping"*—an insight

[4] The waveform displays the amplitude of sound rather than the pitch or frequency.

[5] Sample playback by the hearing first author revealed the sound of heavy rainfall at a similar volume to the thunder.

unavailable to P14. After struggling to connect the waveform to her intuition, P2 was apprehensive about using it again: *"What does that actually mean when it goes up and down? [...] If I don't know the representation that's there, how do I identify [the sound]?"* By contrast, P7 took breakdowns in stride: *"I had never really seen how [the waveform] works. [...] I expected it to be one way, but the waveform showed something completely different. I thought it was a cool experience."*

Subjective Opinion of Sample Quality

When reflecting on our characteristics for determining the samples' quality (Section 4.2.2), participants described uncertainty over how accurately they had replicated their sounds and if they had captured indicative background noise. For replicating sounds, P10 was concerned that her manual reproduction of wind chimes—an otherwise spontaneous sound—was unrealistic: *"It's a different sound. I prayed that in the next few days it was going to be windy enough, [...] [but] it felt like it was cheating."* While P10's residual hearing made her aware of her replicated sound's difference from its real-world counterpart, P12 explained that she did not have the same ability: *"As a deaf person, [...] I'm just relying on my vision and my [other] senses. And so to try and figure out a temperature [alert] or my cat's meow, there are visual indicators, but it's hard to emulate or simulate those [realistically]."*

During the initial session, we defined appropriate background noise as *"the typical sound activity in that location"*, but capturing this proved difficult for many participants during the field study: *"It's hard to differentiate when there's white background noise versus someone talking really quietly in the background, and if that would be interfering [with the sample]"* (P1). As a result, participants said they found it more important to eliminate *all* extraneous noise from their samples than to capture realistic background noise for their context: *"As long as it took full blast on my hearing aids to be able to hear any measure of background noise, I was like, 'you know what, it's fine'"* (P4). When explaining why she chose to *"isolate"* her sounds, P5 said, *"I thought that [doing this] was critical to be able to identify what the sound was and be able to recognize it."* P3, however, was more accepting of the notion of recording unintended sounds: *"My neighbors, they were still making noise; either them talking or their TV or their dog was barking. [...] I can't hear it of course, but my cat was looking around and was drawn to the sound."*

Post Hoc Review Strategies

After recording samples, participants reviewed their samples via audio playback and waveform comparison—with mixed success. With regard to audio playback, five participants said they used their residual hearing to listen back to some or all of their samples (P1, P4, P6, P10, P13). However, P6 included a caveat: *"I would check after recording to make sure I could hear what was going on to the fullest extent that it was possible to do. [...] I do not have the same quality of hearing as a hearing person."* All five participants said they used digital hearing aids or cochlear implants to listen to the audio, which may distort compressed recordings [22]. P10 suggested this issue caused her to avoid using playback: *"The [recorded] sound I heard from my cat was not the sound I hear when my cat's eating. [...] I heard this really loud [*slurping noise*] and I was like, 'Woo! That's a different sound than I am used to.'"* However, the remaining three participants said playback made their review easier: *"I listened to them all eventually with hearing aids. [...] I could just check and go, 'Okay, you know what? That sounds pretty good'"* (P4).

Some participants compared samples to others in their training set to assist with interpreting their contents. For example, P1 judged samples within the same class against each other by listening back in consecutive order, *"[Because] maybe there's something that I didn't catch, even if I think that [sample] sounds good."* Others said that visual comparisons (*e.g.*, flipping between waveforms in the Rev app) were effective for judgments across classes but ineffective for samples within the same class; for example, P9 said he was unsure if he had successfully incorporated the kinds of diversity he had intended for an appliance alert class: *"[The waveforms] didn't really distinguish very well, which made me question, 'Was the dryer beep [that I recorded] really low, medium, and high?'"* A few participants failed to see the utility of the waveform for assessment at all; for example, *"Some of [the waveforms] were skinny and some of them were fat, some of them had patterns and some of them were uniform. [...] [I] was curious about it, but can't say it helped"* (P8). The review was also challenging for P7, and he saw the potential for others to help: *"Relying on hearing people to feed the sound to a machine, [...] that might be better."*

Summary

To interpret their samples, participants used the real-time waveform, listened to post hoc audio playback, and made comparisons to other samples in their training set. However, absent or limited auditory experience led to breakdowns in using the waveforms and uncertainty over how indicative the samples were of real-world sounds.

4.3.5 Ideas for Future Sampling Tools

In the exit interview, we asked participants to brainstorm their ideal sampling tool for building a personalized sound recognizer. We presented a set of design probes [62] (Figure 4.3) to elicit responses to specific features for such a tool. Here, we quantify and qualitatively describe these preferences, which underscore a desire for feedback to (1) better understand the soundscape when recording and (2) provide scaffolding for assessing each sample during post hoc review.

Participants only briefly used the *spectrogram* visualization during the first study session. Nearly all participants were novices and described them as *"confusing"* (P2, P4, P5, P10) and *"overwhelming"* (P6). P12 expressed confusion over the vertical frequency spectrum, noting the difference from her experience with hearing loss testing: *"I was thinking the lower [volume] would be on the top. [...] For auditory tests, [...] on the right-hand side is where you see [high frequency] on the audiogram.*[6] *[...] It threw me off."* Only two of 14 participants chose to include spectrograms in their ideal tool: P14 due to using it extensively in coursework (*"I'm able to notice more of the texture of sound"*), and P2, who thought it could tell her when *"three or four different sounds are happening because I saw three or four different colors."*[7]

Reinforcing their positive experience with the *waveform*, 11 of 14 participants chose to include it in their ideal tool, calling it *"helpful"* (P7, 12) and appreciating its simplified temporal and volume information. P7 said that without the waveform, *"I think that the background noise would have interfered, because [...] I'm not able to hear [that]."* A real-time *volume monitor* (Figure 4.3a) was only chosen by eight participants, and most preferred Rev's real-time waveform instead. However, P5 thought it could *"let me know that something was coming"* after failing to record passing vehicles. Five participants wanted to include the waveform with the spectrogram for *"more information"* (P8) when needed: *"[The waveform] has very concise information of what's actually necessary, the spectrogram captures everything in the environment"* (P1). Notably, P10 rejected

[6] An audiogram displays the results of a pure-tone hearing test, the gold standard measure of hearing loss [140] using a 2D frequency-volume graph. Frequency increases to the right on an audiogram, while volume (as dB loss) decreases moving upward. For more information, see: `https://www.asha.org/public/hearing/audiogram/`

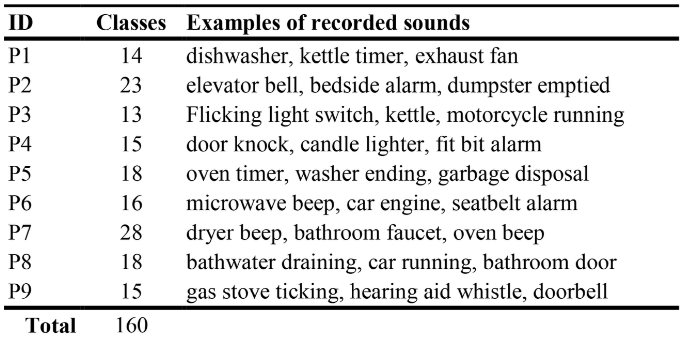
[7] Color is used on the spectrogram to show amplitude rather than distinguish sounds.

both visualizations, trusting herself to hear the soundscape instead: *"I knew that it was recording [correctly]. [...] I could hear with my hearing aids, even though the sound was distorted."*

Many participants desired enhanced awareness of co-occurring sounds; 12 of 14 wanted a real-time *overlapping sound monitor* (Figure 4.3c) and P12 explained, *"I can't hear things happening at the same time [...] I don't know if it's the cat meowing or the TV blaring or the washing machine has stopped."* Ten participants wanted a real-time *background noise monitor* (4.3b) to show the ambient noise level separate from any unique sounds, while nine opted for a post hoc *background noise removal* option to remove any undesired artifacts from their samples. However, P2 worried processing could also remove the personal elements of her samples: *"Squeaky clean—it's not normal."*

Participants desired clearer feedback on the contents of their samples after struggling with interpretation in situ. The post hoc *quality rating* (Figure 4.3f) was selected by 12 participants, although they said it lacked utility out of context: *"If the problem [...] was detected, then I'd have to figure out how to resolve it"* (P3). P5 favored automatic assessment, saying simply, *"I don't trust myself when it comes to the sound."* All participants selected post hoc *background noise feedback* (4.3d) and 11 added *overlapping sound feedback* (4.3e), hoping both could alleviate *"doubt"* (P1) over the samples' contents and with *"determining whether to re-record"* (P4).

Participants presented their own ideas for features, and the most prevalent was a *"hypothesis of what the machine is hearing"* (P6) that would be drawn from *"a big dictionary of sounds"* (P13). P9 wondered if he could guide the hypothesis: *"I type in, 'I'm going to be recording a refrigerator beep.' Then the system would know I'm looking for beeps. It would help [the system] in the process of elimination."* P2 explained her desire to know more about class similarity: *"I'm curious what made the sound I chose [...] different from something else. [...] Like two different doors: do they go with 'those two doors sound the same,' or are they different?"* Drawing from in situ strategies, a few participants proposed using visualizations generated from larger sound libraries to guide their expectations of their own samples: *"Being able to see what birds chirping might look like on [a waveform] [...] and then when I record it, making sure that [my] waveform is matching"* (P7). Finally, P1 wanted to make the sampling process more closely resemble the end-to-end training of Google's Teachable Machine: *"At the end of it, you could actually try to repeat a sound, and it would capture, like, '90*



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	

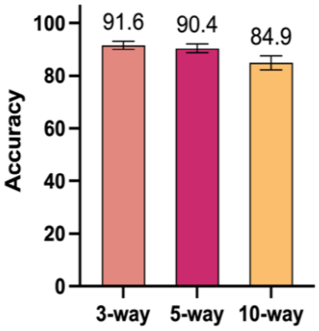


Figure 4.5: (a) DHH participants' recorded sound class counts with examples. Note that many of these classes are highly specific to participants' use cases (*e.g.*, flicking light switch, hearing aid whistle) and thus, require model personalization. (b) ProtoSound's average accuracy for 3-class, 5-class, and 10-class evaluations on DHH participants' recorded sounds.

percent crumpling paper', '30 percent clapping'. Seeing those kinds of feedback there, [I] was like, 'Oh, this is actually recognizing, indicating, positing the sound.'

Summary

Participants responded positively to features that would inform of soundscape activity, especially to distinguish when sounds overlap or interfere with the sound of interest. In addition, they requested support in determining how the machine would interpret each sample compared to a larger training set.

4.3.6 Analysis of Participants' Audio Samples

While not included in the original publication [47], we also conducted a follow-up analysis of participants' audio samples from the field study to support Jain *et al.*'s ProtoSound work (published in [67], Sec. 6). The analysis demonstrates the potential for models created with the ProtoSound architecture to achieve high accuracy when trained with a dataset recorded by DHH users and to successfully accommodate a reasonable variety of their desired sounds.

To construct a dataset relevant to our analysis, we chose participants who 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 over one second were removed. Class counts per participant, including example

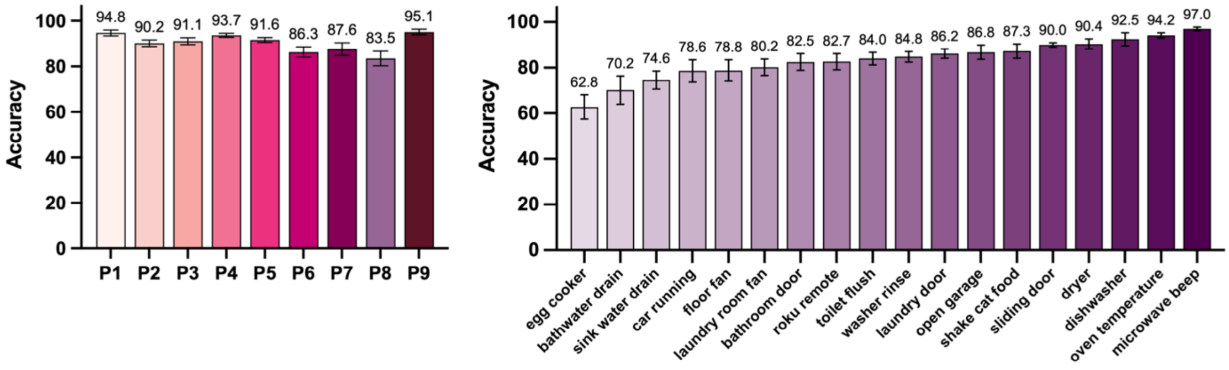


Figure 4.6: (a) Average accuracy per DHH participant for the 5-way setting. (b) Accuracy per-class for the lowest performing case: P8. Note that 'egg cooker' performed poorly due to user recording errors in some samples (missing sound). 'Bathwater drain' and 'sink water drain' performed poorly since they were very similar in sound and confused with each other.

classes, are shown in Figure 4.5 a. 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.

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 4.5 b for results. For the 5-way setting—the most desired by DHH people, according to our survey [67]—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 4.6. 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).

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%, and 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.

4.4 Discussion

This study confirms the potential for non-expert DHH users to train personalizable sound recognizers (as identified in past work [107]) and advances understanding of: (1) how non-expert DHH users approach in situ recording tasks to create a sound recognizer training set, (2) practical challenges that they may face when recording a variety of real-world sounds, and (3) sense-making strategies that they use to interpret audio data in this context. Here, we discuss the implications of our findings, opportunities for future work, and the limitations of this chapter.

4.4.1 Technical Implementation

Our work fits within a supervised ML context where DHH end users capture and label audio samples to train a sound recognition model for custom sound classes. Regardless of whether the system is ultimately implemented using a single batch training process (*e.g.*, one-time collection of a set of samples to train the model) or a more interactive ML approach (*e.g.*, iterative training and refining of the model), DHH users will need to capture audio samples.

Most sound recognizers both for general tasks (*e.g.*, [17, 87]) and specifically for DHH users [71, 72, 127] adapt deep learning approaches from computer vision—such as *VGG* [128] or *ResNet* [54]—and use transfer learning [135] to train on a large dataset of sounds such as AudioSet [44], FreeSound [40], or field recordings. For example, *SoundWatch* [72] is a *VGG*-based smartwatch app that supports 19 sound classes, was tested with DHH participants, and is now available as an open-source application [93]. An initial personalizing step might be to adapt such a model to enable fine-tuning [150] for an individual user's sounds (*e.g.*, a generic dog *vs.* my dog). While fine-tuning has shown promise for personalizing activity recognition models [2], the supervised approach is data intensive, and some DHH users may be uninterested in building a large training set themselves [107]. Meta-learning [39] can reduce the necessary data by generalizing information about several related tasks to the new task and may realize a few-shot learning approach to sound classification, allowing DHH end-users to train custom sound classes with just a few samples of their own—a task that all of the participants in our study deemed reasonable.

While the approaches outlined above can allow for models trained from an end user's samples, our results suggest that systems intended for DHH users should also allow for other data sources. For example, participants recorded few samples of urgent sounds during our study (Figure 4.4) despite this category being the most widely requested sound awareness tool by DHH users [13, 37]. Our participants explained that although they desired more samples of urgent sounds, many of these were infrequent and uncontrollable (*e.g.*, gunshots, building fire alarms). Unlike DHH users who depend on visual cues, hearing users may be able to catch part of a prolonged sound with no visual cues—such as an approaching siren—but they would likely face similar challenges for shorter, spontaneous sounds. To account for this, systems should be designed to support user-provided audio in addition to samples from sound libraries, such as Nakao *et al.*'s [107] design that allows the choice between a recording tool or AudioSet [44] search.

DHH users who desire a personalized model but feel unqualified to record samples, such as P7, provide impetus to explore additional techniques. With reinforcement learning, the system can be incentivized to adjust its behavior based on positive and negative feedback, allowing users to guide the model to better fit their needs [81]. For example, a recognizer could prompt the user for post hoc assessment of each recognized sound and refine itself over time—an approach that has shown promise with deep-leaning models for automatic speech recognition [116]. However, while a DHH user may feel comfortable assessing this output in a familiar location (*e.g.*, their kitchen), they may find this task challenging in unpredictable contexts.

This reinforcement learning approach raises the question of how DHH users can assess a recognizer's output when they themselves are unsure about a sound. Combining multiple models may support a comparative evaluation of the sound; a similar technique was leveraged by our participants for interpreting their samples. To support evaluation for batch learning personalization, designers can display the custom model's output next to output from a pre-trained model supporting broad sound categories (*e.g.*, [72]). Several of our participants even requested a *"hypothesis"* (P6) after recording each sample, but we found their interest in the model's state—while present—was secondary to their overall uncertainty about the sound itself. Approaches leveraging multiple models have also been used for semantic data representation (*e.g.*, navigating a large audio dataset [63]); a pre-trained model could additionally provide DHH users with a speculative classification of each sample to compare with its labeled sound class.

Our study did not involve using a human-in-the-loop system past the brief demonstration of Google's Teachable Machine system, but our results motivate future explorations of these systems with DHH users. For example, real-time cause-and-effect feedback afforded by human-in-the-loop systems (*e.g.*, [35]) could provide insight into how an individual's samples shape the model, while user-defined decision boundaries (*e.g.*, [4]) could allow DHH users to tolerate errors for less critical sounds (*e.g.*, birds) but not others (*e.g.*, alarms). However, most deep learning algorithms currently underpinning sound recognizers do not support direct interaction (*e.g.*, adjusting parameters) [30], and future work intending to leverage the benefits of human-in-the-loop systems for DHH users should explore alternatives.

4.4.2 Design Suggestions: Instruction, Visualization, and Feedback

Our study uncovered unique pitfalls that non-expert DHH users may encounter while recording samples to train a personalizable system. This section proposes possible solutions to these challenges through specialized instruction, enhanced audio visualization, and additional feedback to aid review.

Informed by our participants' experiences, we synthesize four sound dimensions that designers of sound sampling tools should consider when supporting DHH users: (1) *Volume & frequency:* How loud is the sound? What range of frequencies are in the sound? Are these properties stable (fire alarm) or shifting (baby crying)? (2) *Length & continuity:* What is the duration of the sound? Is the sound continuous (a fan) or disjoint (typing on a keyboard)? (3) *Locus of control:* What is the user's role in reproducing the sound, from direct (clapping) to indirect (pet sounds) to none (emergency sirens)? (4) *Consistency:* How varied is the real-world population of the sound, from uniform (phone rings) to moderate (musical instruments) to highly diverse (television)? Each DHH user's ability to record a given sound will also depend on personal and contextual factors such as residual hearing ability (*e.g.*, use of cochlear implants *vs.* no device), lifetime experience with sound (congenital *vs.* post-lingual hearing loss), and recording location (*e.g.*, a quiet home *vs.* a busy park).

Prior work shows that non-expert users often misconceptualize how ML systems work [82, 139], and instructional scaffolding can improve their understanding and satisfaction with personalized ML tools [83]. Prior work on non-expert ML use often provides scaffolding guidelines; for example, Yang *et al.* [148] suggests "test-driven machine teaching" to guide non-experts through training via real-world test cases. However, to meet DHH users' needs when recording sounds for ML, we suggest that audiological topics be included in this scaffolding. First, to support DHH users' conception of the system's decision-making process, provide an explanation for the sound features used as input to the model (*e.g.*, two-dimensional spectrograms [55]) and show variations of these features in samples of the same sound. For example, several of our participants believed all samples for a class should be recorded at similar volumes, which may not be required for an ML system yet complicated their experience. Second, to support DHH users' understanding of a model's decision boundaries, provide an overview of common sounds and their distance from one another on the model's decision axis. Although a machine processes sounds differently than a human, a hearing user may be able to identify relative differences of consequence to a machine (*e.g.*, similar appliance beeps). A DHH user, on the other hand, who cannot hear that sound at all, may be forced to guess or *"imagine"* (P9) these differences instead.

A user's ability to interpret data is essential for training a personalized ML system. Hearing users can assess the contents of their sound samples both by listening to the soundscape while recording and by playing the audio back afterward, but equivalent techniques are not reliably available to DHH users—even those who used residual hearing in our study. Participants liked waveforms for recording in a lab setting [13], and most of our participants agreed the Rev app's [119] waveform visualization was crucial for recording in situ. However, breakdowns in our participants' waveform sense-making highlight the potential for more intuitive visualizations to DHH users and informative about the recognition model. For example, during limited use of spectrograms, most participants found them difficult to interpret—reflecting the known difficulties both hearing [18, 63] and DHH [99] novices can have with spectrograms. Yet these visualizations are shown to be powerful for experienced users [134]—including DHH ones, such as P14—and many sound recognizers extract features directly from spectrograms [55]. Interpretation of a sample's spectrogram on its own may be naive, but it may be useful to compare spectrograms across samples, a strategy our participants used with waveforms. Designers could also investigate other time-frequency visualizations to inform DHH users in this context, such as correlograms and pitchograms [20], or explore new visualizations based on audiograms (2D frequency-volume graphs that are widely used in hearing loss testing [140] and referenced by our participants).

While sound visualizations can help to reveal the full soundscape to DHH users, our participants were also enthusiastic about high-level feedback for audiological information. Because many DHH users cannot hear the real-world version of the sound they are recording, they may also be unable to determine how closely a sample fits within the broader population of that sound. A 2D feature-embedding generated from the data [63, 113] can provide a sense of the diversity of the data in question (*e.g.*, if a class clusters together, if a sample is far from its counterparts), but DHH users may have a more significant issue in determining *why* a sample is different from others. Many participants were also uncertain about co-occurring or overlapping sounds when recording, while others desired insight into the ambient soundscape (*i.e.*, background noise). While these artifacts alone may not impact a sample's quality as training data—processing algorithms such as independent component analysis [94] may separate sources or negate the impact of ambient sound [26]—additional feedback that informs DHH users about these artifacts may greatly enhance a DHH user's insight into the contents of the sample.

4.4.3 Sociocultural Implications

Researchers should also carefully consider how to create tools for interested DHH users while not inscribing audist beliefs. We encountered a diversity of perspectives in our study that reflects the wide-ranging needs and preferences of the DHH community. While we did not encounter opposition to our envisioned recognizer in our study, we do not assume it is universally desired: other DHH people may feel negatively towards this technology, especially those who identify as Deaf and as part of Deaf culture [10]. However, our study reiterates prior work showing the strong situational value that a sound recognizer can provide for some DHH people [13, 71, 72, 127], and it is possible that some DHH users may desire enhanced awareness of a few highly situational sounds while otherwise avoiding the hearing world. A personalizable sound recognizer that could be constrained to detect only a small subset of sounds (*e.g.*, a child's cry) may provide essential support while preserving a user's cultural preferences. In addition, although we designed our study to believe that a system should support independent personalization as a baseline for DHH users, our findings suggest some users may still feel unqualified for this task. Several of our participants enlisted support from hearing and DHH family and friends when recording, which, in combination with collaborative benefits seen in a workshop setting [107], suggests that *interdependent* usage may be natural to some DHH users.

4.4.4 Limitations

First, we focused on DHH users' needs during data collection and review, and we did not examine other stages of training a sound recognizer following the initial session. While Nakao *et al.* [107] provides an analysis of DHH users' engagement with an IML system, training a working model from our participants' samples and allowing them to engage with it *in situ* may have provided greater insight toward their conception of this space. Second, conducting this study during the COVID-19 pandemic limited our participants to those with high-speed internet access and time to join a research study amidst social, health, and economic uncertainty. Finally, the remote nature of this study and COVID restrictions limited participants' in situ recording contexts and our ability to directly observe recording activity, and our request that they do not ask anyone for help when recording sounds reduced the realism of the in situ scenario. For a complete understanding of practical recording, future work should study recording experiences across various locations and allow users to solicit feedback from hearing people.

4.5 Chapter Summary

Prior evaluations of sound recognition tools with DHH users highlight a desire for custom sound classes from user-provided recordings; however, prior work has not explored how these users record and engage with real-world audio data. This chapter detailed one-week field study on the experiences of 14 DHH participants as they collected and reflected on audio data for the purpose of training a personalized sound recognition model. Our findings highlight the practical challenges of capturing spontaneous, infrequent, or visually obscured sounds, and the sense-making strategies used to interpret audio data when auditory feedback is unreliable. While participants appreciated the support of waveform visualizations, they desired additional feedback to monitor ambient soundscape activity, assess their recordings' representativeness of real-world sounds, and better understand the composition of their dataset. As detailed in the following chapter, the insights from this study were integrated throughout the design of the SPECTRA prototype, including its tutorial, real-time visualizations, sound library, and clustering for high-level insight into the dataset.

Chapter 5

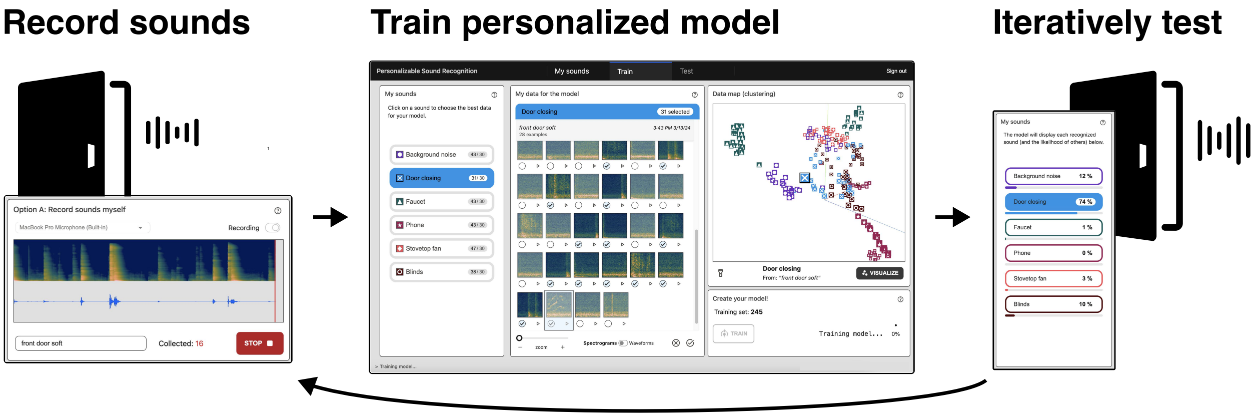
Evaluating an Accessible Sound Recognition System[1]

5.1 Introduction

The sounds recorded by participants in the previous chapter's field study are emblematic of DHH individuals' wide-ranging sound interests. In general, users are interested in sound information to augment personal safety (*e.g.,* footsteps), social engagement (*e.g.,* nearby voices), and everyday tasks (*e.g.,* monitoring home appliances) [13, 37, 71]. To meet users' sound interests, sound recognition tools have proliferated, both in the research literature (*e.g.*, [72, 102, 127]) and in commercial applications—for example, Android and iOS smartphones support sound recognition for common sounds like doorbells, running water, and dog barks.

Despite these advances, DHH users have expressed a need for improved sound recognition accuracy and support for a wider range of sound categories [60, 67]. As demonstrated in Chapter 3, one challenge is that the value of sound information is highly contextual: hearing identity [37], social context [13], physical location [46], and individual preferences [37] can all influence how a DHH user may benefit from

[1] This chapter includes materials originally found in [48] on the design and evaluation of SPECTRA, an interactive prototype for the accessible creation of sound recognition models.



Click on a sound to choose the best data for your model.

The model will display each recognized sound (and the likelihood of others) below.

Figure 5.1: Overview of the SPECTRA pipeline, showing data collection, selecting a training dataset, and testing. The design includes key elements to support the needs of DHH users throughout this process, including spectrogram and waveform visualizations, rich descriptions, and a three-dimensional feature-embedding chart.

sound information. These contextual differences could be addressed by empowering DHH individuals to personalize a sound recognition model themselves [13, 47] or fine-tune a model to their local soundscape [71]. Personalization also has the potential to improve accuracy by providing examples specific to an individual user's context, such as the sound of *their* appliance or alarm rather than a generic one.

However, an open question lies in how to effectively support a DHH user—who does not have full access to a sound themselves—in capturing and selecting suitable audio data to train a machine learning (ML) model [47]. Android and iOS recently introduced the ability to add custom sound categories [51] or tune the model for specific alarms and appliances [7], respectively, through a brief recording process. While this approach is simple, allowing users with or without ML expertise to engage more directly with the machine learning pipeline through an interactive machine learning (IML) approach provides a sense of transparency and control [29], which can positively impact trust, satisfaction, and long-term use [3, 83]. Prior work has begun to investigate the potential of IML for sound recognition systems for DHH users, but, in one case, did not provide non-auditory feedback about the data to DHH participants [107] and in a second, only focused on data collection within the ML pipeline [47]—not including model training, evaluation and iteration.

In this paper, we introduce and explore *SPECTRA—Sound Processing and Enhanced Custom Training for Recognition Assistance*—an interactive pipeline for the accessible creation of personalized sound recognition models. Our human-centered approach merges IML design guidelines [30, 117, 148] and the needs of DHH users [47, 107] to train a personalized sound recognizer via IML. To evaluate how SPECTRA supports DHH users in engaging in IML, we recruited 12 DHH participants who each trained a personalized sound recognition model for their home soundscape. We examined how spectrogram and waveform visualizations, interactive clustering, and rich text annotations can support DHH users across an interactive training cycle, and how experience with these mechanisms may shape performance expectations, technical understanding, and confidence. We also investigated the impact of the experience on participants' perceptions of and attitudes toward personalizable sound recognition models.

Our findings reveal new insights to support DHH users in personalizing sound recognition models, including demonstrating how interactive data clustering, in combination with waveforms, can enhance DHH users' understanding of audio data, identification of outliers, and refinement of training datasets. We show the value of non-auditory data representations for DHH users at different stages of an end-to-end training cycle (data collection, training data selection, model testing) and explain how they incorporate this information into their reasoning about sound models. Our results also reaffirm prior work showing DHH users' preference for waveforms when recording [47]—while expanding on their value when selecting training data and testing—and show how users' training strategies develop through use [107]. Finally, we provide insights into DHH users' experiences and perspectives on personalizing a sound recognition model.

In summary, our work contributes: (1) SPECTRA; a novel, end-to-end pipeline to support DHH users with capturing sound examples, curating a training dataset, and testing the models they create; (2) results from a qualitative evaluation to understand the system's benefits and obstacles for DHH users, including its impact on their conceptualizations of sound recognition models; and (3) design considerations and recommendations for future systems that meet the needs of DHH users during interactive training tasks.

5.2 SPECTRA: A DHH-Centered Pipeline for Personalized Sound Models

To investigate how visualization techniques can support DHH users in personalizing their own sound recognition models, we built SPECTRA (Sound Processing and Enhanced Custom Training for Recognition Assistance), a prototype IML web application. The pipeline's design was informed by related sound awareness literature [13, 47, 67, 107] and guidelines for human-centered ML systems [30, 117, 122, 148]. SPECTRA

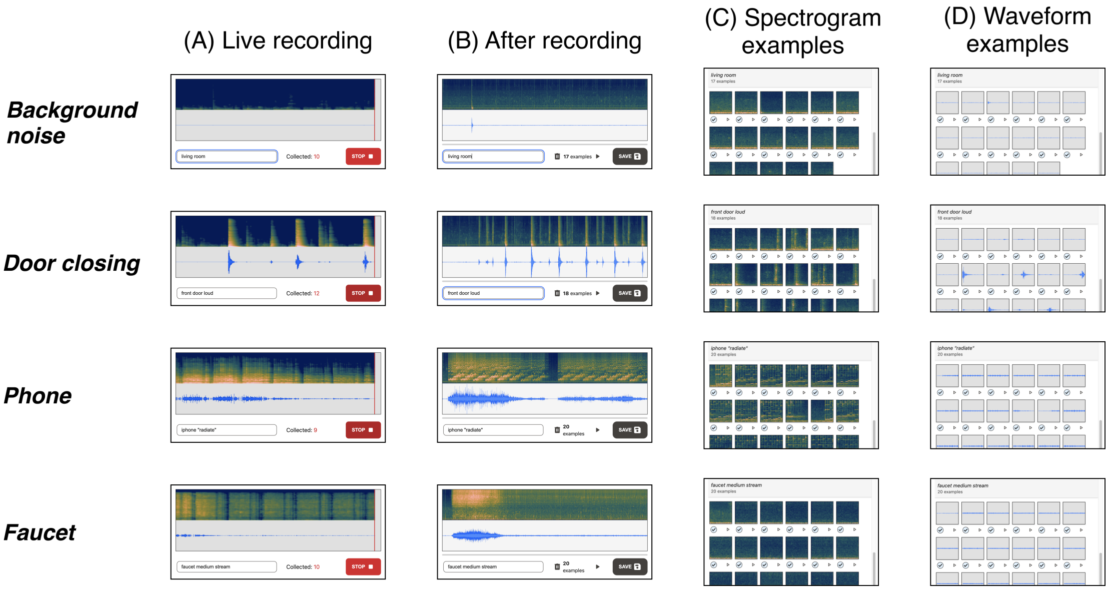
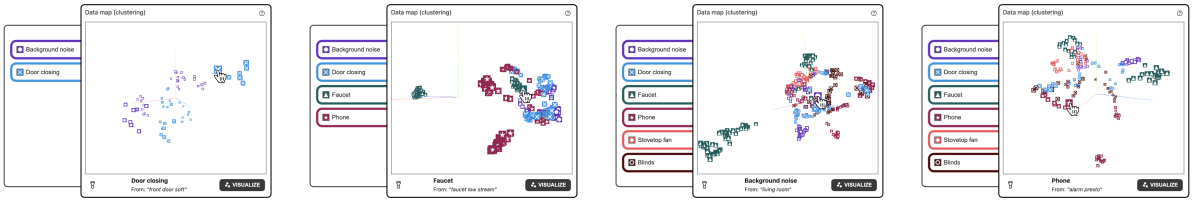


Figure 5.2: Spectrogram and waveform visualizations for four different sounds within the SPECTRA workflow. (A) While recording, a stacked spectrogram and waveform display streaming audio input over a 10-second window. (B) After recording, the stacked visualization shows the full duration of the captured audio. When selecting training data, each recording is segmented at 1-second intervals, which users can choose to display as (C) spectrogram or (D) waveform icons.

has a three-step workflow: users first generate a training dataset, then edit the training set and generate a model before testing the model's real-time sound recognition capacity. In this section we describe the implementation of each step and outline the pipeline workflow and functionality.

5.2.1 Spectrogram and Waveform Feedback

SPECTRA uses high-fidelity waveform and spectrogram visualizations to convey audio data to DHH users (Figure 5.2). Waveforms show the amplitude—or loudness—of audio over time and are common in audio recording, editing, and playback software. DHH participants in prior work reported waveforms were intuitive and useful for capturing audio examples, though the amplitudinal feedback alone was inadequate for verifying the recordings' quality (*e.g.*, no co-occurring sounds) [13, 47]. They further requested that the visualizations remain active before and after recording to monitor the ambient soundscape, while audio playback helped those with residual hearing to analyze waveforms with unclear meanings [47]—both of which are included in SPECTRA.



(A) 2 sounds **(B) 4 sounds** **(C) 6 sounds** **(D) 6 sounds, after curation**

Figure 5.3: Clustering visualizations within SPECTRA, showing the structure of unedited audio datasets with (A) two sounds, (B) four sounds, and (C) six sounds. As the size/variety of the dataset increases, so does visual complexity. (D) Curating the raw six-sound dataset (*i.e.*, removing mislabeled examples) shows greater separation for many clusters, though some overlap persists. Note: Hovering over a data point displays the example's label and the annotation (metadata) of its parent recording.

Spectrogram visualizations offer greater information throughput by visualizing both amplitude and frequency and are commonly used for discriminating noises in environmental soundscapes (*e.g.*, [25]). While spectrograms can be powerful data interpretation tools for experienced users [18], DHH participants had a mixed response following brief use in a lab setting [47]. Models trained with SPECTRA—like many sound recognition models (*e.g.*, [67, 127])—take Mel spectral features (*i.e.*, frequencies bucketed to approximate human hearing) as input, meaning that users viewing spectrograms see the same audio properties the model uses to make decisions.[2] We include both visualizations to cater to DHH individuals' diverse preferences and learning styles and explore their effect on users' decision-making about a training dataset.

5.2.2 Interactive Data Clustering

SPECTRA includes a three-dimensional data clustering visualization to help DHH users understand and refine an audio dataset (Figure 5.3). We draw from prior work on interactive data clustering [9], including sound clustering with hearing users [63], to address the unique challenges DHH users face in an iterative training process. Goodman *et al.* found DHH individuals had issues with discerning variations in sounds (*e.g.*, porcelain *vs.* metal faucets) and anticipating how audible differences will affect model performance [47].[3] SPECTRA's clustering visualization complements the waveform and spectrogram displays—which show individual audio instances—by rendering the structure and diversity of the broader dataset. We employ

[2] When displaying Mel spectrograms, SPECTRA converts amplitude values to a logarithmic dB scale; this "log-Mel" scaling more closely aligns with how humans perceive sound.

[3] Although neural networks "hear" sounds differently from humans, hearing users can use audible differences to make a relative estimation of potential issues within a model (*e.g.*, garbage disposal and coffee grinder); DHH users may lack this ability.

UMAP dimensionality reduction [101] to project high-dimensional spectral audio features into an embedding space, where similar examples colocate while distinct examples separate.

UMAP is noted for its speed and preservation of datasets' global structure, which may help DHH users to better understand and make informed decisions about their training datasets. During development, we determined that, though more complex, three-dimensional embedding space allowed for greater visual discrimination between clusters. With SPECTRA's 3D visualization, users can rotate and zoom to explore the clusters and identify outliers, ambiguous examples, or underrepresented classes. For instance, a cluster of data points labeled as "dog bark" that appears distant from other dog bark clusters might prompt investigation into unusual background noise or varying bark types. While not directly visualizing model parameters, clustering visualizations may guide how users choose to refine their training dataset, such as by removing outlier examples (if determined to be unrepresentative or mislabeled), merging or splitting classes, or collecting additional data. After updating the training dataset and regenerating the clustering visualizations, users can see how their changes have impacted the separation of their classes. Thus, SPECTRA provides users with an ongoing and evolving representation of how differentiable their data is and better guides users through the iterative training process.

5.2.3 Rich Data Annotations

During data collection, SPECTRA allows users to annotate their recordings with textual descriptions, capturing contextual details (*e.g.*, water running from bathroom *vs.* kitchen sink). Annotations serve as a form of semantic metadata for users, separate from model labels. Many real-world sounds vary depending on their source, production method, or environment—a challenge when personalizing sound recognition models, where users need to provide a representative dataset for the model to generalize to their environment. DHH users in prior work questioned the meaning of differences among their recordings and the impact of sound variations on their models' performance [47]. SPECTRA encourages users to identify and capture variations of each sound (*e.g.*, faucet → stream, drip), drawing from the concept decomposition process of the interactive machine teaching paradigm [117, 142]. Annotations allow DHH users to document their domain expertise not readily apparent in SPECTRA's visual feedback alone, such as to clarify subtle differences in waveforms/spectrograms or to support reasoning about loose or separated clusters with the same sound label. Annotations are displayed alongside SPECTRA's visualizations during data selection, serving as a memory aid and reasoning tool to make sound data more understandable to DHH users.

5.2.4 SPECTRA Implementation and Workflow

We built SPECTRA using *Node*[4] and *Svelte*[5] from a fork of *Marcelle.js*[6], an open-source toolkit for creating ML workflows and interfaces [41]. To enable transfer learning from a pre-trained sound model, SPECTRA is powered by the *Speech Commands API*[7] from *Tensorflow.js* [136], which employs a convolutional neural network (CNN) pre-trained on the Speech Commands dataset (50K examples from 20 classes) [144]. CNNs are commonly used in sound recognition due to their ability to learn complex patterns in audio data [55]. SPECTRA uses the API's transfer learning functionality—where a pre-trained model is re-used as a feature extractor for new classes, reducing training time and resources—to apply the speech-trained base model to the environmental sound domain. We chose this library for its ease of development, rapid prototyping capabilities, suitability for in-browser use, and lightweight package. These allowed us to focus on the interactive aspects of the system and study how DHH users engage with the IML workflow.

To custom train the sound recognizer, users capture continuous audio recordings via the web browser's built-in *Web Audio API*. To match our classification model's input features (1-second Mel spectrograms with a 43×232 shape size), users' recordings are segmented at 1-second intervals, converted to spectrograms using the Short-time Fourier transform (STFT), then converted from the linear frequency scale to the logarithmic Mel scale. These Mel spectrograms are presented to users as "examples", and the user can select specific segments to include as training data.

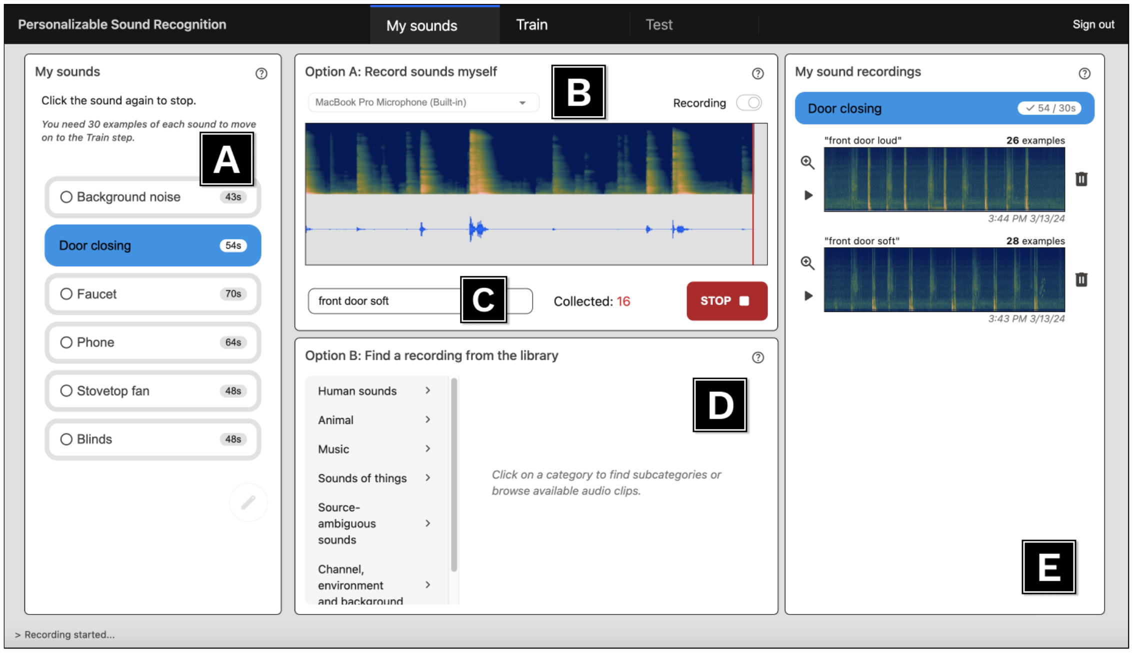
While prior work has explored mobile devices for recording audio for personalizable sound recognition [13, 47], we focus on the entire IML workflow and thus designed SPECTRA for laptop/desktop screens. We do not assume this is an ideal format for end-users; instead, we leverage the large screen size to present multiple high-fidelity visualizations in tandem (waveform, spectrogram, and/or clustering) and learn about salient information to assist DHH users when personalizing a sound recognizer. SPECTRA's UI is organized

[4] Version 18.12.1. https://nodejs.org/

[5] Version 3.48.0. https://github.com/sveltejs/svelte

[6] Version 0.6.0. https://github.com/marcellejs/marcelle

[7] Version 0.5.4. https://github.com/tensorflow/tfjs-models/blob/master/speech-commands/



Click the sound again to stop.

You need 30 examples of each sound to move on to the Train step.

front door soft

Click on a category to find subcategories or browse available audio clips.

"front door loud" 26 examples 3:44 PM 3/13/24

"front door soft" 28 examples 3:43 PM 3/13/24

> Recording started...

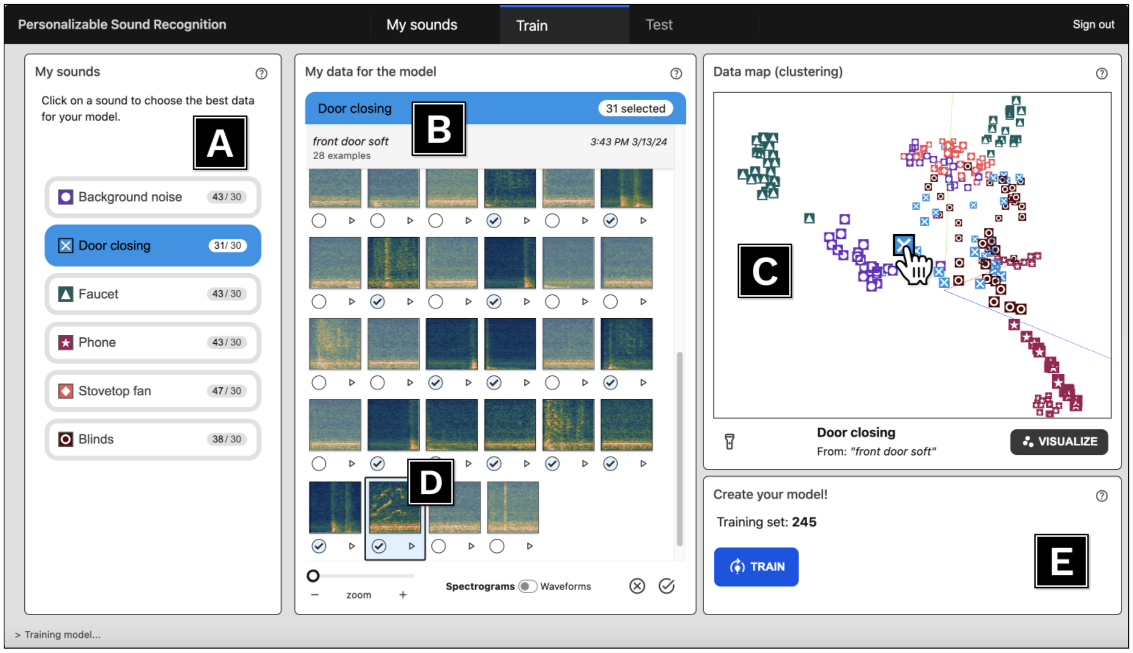
Figure 5.4: The "My Sounds" page, used for data collection. (A) Users select a sound class to begin collection. (B) They can record with a live waveform and spectrogram of their device's microphone input, and (C) annotate the recording with relevant contextual details. (D) Alternatively, users can select pre-recorded examples from a categorized library of video clips. (E) Users can review collected recordings for the selected class.

into three tabs ("pages") corresponding to different stages of the IML workflow [30]: (1) planning and data collection; (2) data curation and model training; and (3) iterative model testing.

Planning and Data Collection

SPECTRA users start at the "My sounds" page (Figure 5.4), which aligns with the planning and data collection stages of a typical interactive ML workflow (*e.g.*, [30]). Users first define and create placeholder classes for desired recognizable sounds in the "My sounds" panel (*e.g.*, "my dog barking" or "stovetop fan"). SPECTRA currently supports adding up to 10 distinct classes of sounds. The "My sounds" panel (Figure 5.4a) is available across all three pages on the left sidebar. The user then selects a specific sound class to initiate data collection, which activates the center and right-side UI panels.

To collect data, users navigate to the center panel (Figure 5.4b) and click the "Start listening" toggle. Live microphone data is then visualized via the waveform and spectrogram visualizations (but not yet recorded).



Click on a sound to choose the best data for your model.

Background noise	43 / 30
Door closing	31 / 30
Faucet	43 / 30
Phone	43 / 30
Stovetop fan	47 / 30
Blinds	38 / 30

Door closing 31 selected

front door soft 3:43 PM 3/13/24
28 examples

Door closing
From: "front door soft"

Create your model!
Training set: **245**

> Training model...

Figure 5.5: The "Train" page, used for training data curation. (A) Users select a sound class to filter their sampling set. (B) Examples are shown as 1-second spectrograms, which can be toggled for inclusion in the training dataset. (C) Users generate a three-dimensional clustering of the selected training data and can rotate or zoom to inspect the clusterings. Hovering on a point will show its label and text annotation; (D) clicking on it will highlight the example in the data selection panel. (E) Users train a model with the selected training dataset.

After resolving any potential unwanted background noise, the user can then select the "Record" and "Stop" buttons to collect data. While recording, SPECTRA shows users a running count of the collected examples (*i.e.*, number of 1-second spectrogram increments). Users can add text annotations before or after recording to note additional information, such as sound variations (*e.g.*, bathroom *vs.* kitchen sink) or unplanned sound activity (*e.g.*, a cat meowing mid-recording) (5.4c). Alternatively, for hard-to-record or unavailable sounds, (*e.g.*, sirens), users can import existing recordings from the *AudioSet* library of categorized YouTube clips [44] (Figure 5.4d). If they are satisfied with the recording, users can save it to their sampling set, which shows up on the right-side pane under "My sound recordings" (Figure 5.4e). Before users can move on to SPECTRA's "Train" page, they must collect at least 30 examples of each sound (encouraged to be varied across a few 5-10 second recordings)—a threshold selected to ensure sufficient data for model training without overburdening users.

Data Curation and Model Training

In the second stage, users navigate to the ”Train” page, where they review their data, refine the training dataset, and train a model (Figure 5.5). On this page, when users select a sound class from the “My sounds” panel (Figure 5.5a), the center panel populates with one-second examples of that sound (Figure 5.5b). The examples, grouped by recording, are presented as spectrograms, with the option to switch to waveform visualizations. Users can select which examples to include or exclude in the training dataset, with all examples included by default.

SPECTRA’s emphasis on data iteration aligns with practices observed among expert ML practitioners, who typically prioritize dataset refinement over changes to the models themselves [57]. To support users’ understanding of their dataset, SPECTRA includes an interactive data clustering panel (“Data Map”; Figure 5.5c). As they filter out low-quality or unrepresentative examples, users can generate new embeddings of the updated dataset in latent space to monitor how the overall structure changes. The clustering visualization uses UMAP for dimensionality reduction [101] (enabled by *UMAP-js*[8]) of the high-dimensional Mel spectrograms as training features. For simplicity, we chose pre-set UMAP parameters[9] after testing with several audio datasets. These 43×232 arrays are reduced to a 1×3 size and plotted in 3D space with *ScatterGL*[10], using a unique symbol to represent each sound class in the embedding space (located next to that class in the “My Sounds” panel). While embeddings generated from SPECTRA’s base model’s feature space may provide better scalability and align more with perceptual similarity, our dimension-reduction of the Mel spectrograms provided a fast and light-weight method that was acceptable for the constraints of our evaluation—serving as a design probe into how DHH users can be supported to reason about algorithmic interpretations of sound.

Users explore the clustering visualization by rotating (click and drag) or zooming (scroll), and they can select an individual point to highlight its corresponding example in the selection panel for further inspection (5.5d). Users can iteratively adjust the training set and observe its effect on clustering, moving

[8] Version 1.3.3. https://github.com/PAIR-code/umap-js

[9] *UMAP-js* parameters: `nComponents` = 3, `nEpochs` = 500, `nNeighbors` = 20, `minDist` = 0.1, `spread` = 1.0, `supervised` = `false`

[10] Version 0.0.13. https://github.com/PAIR-code/scatter-gl

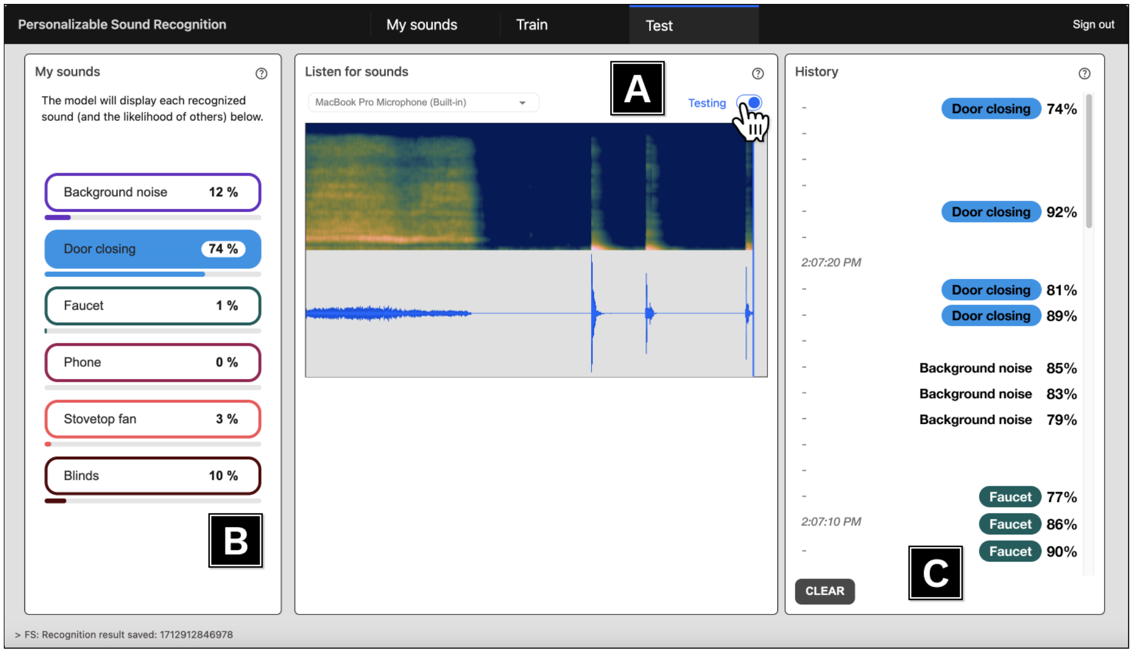


Figure 5.6: The "Test" page, used for practical assessment. (A) Users toggle their microphone "on" to begin streaming recognition. (B) Users produce sounds in the environment and observe the model's predictions as confidence scores (with a bar graph and percentage). (C) A vertical timeline also shows the highest-scored sound at each second of the previous two minutes.

toward more clearly differentiated sound classes. Once satisfied with the refined training set, clicking the "Train" button creates a new model, which users assess on the third page.

Model Testing

In the third and final stage, users navigate to the "Test" page, where they can assess the practical performance of their personalized model's latest iteration (Figure 5.6). Users activate a toggle switch to initiate streaming recognition (Figure 5.5a) and can then produce sounds to evaluate the model's predictions in a live environment along with both waveform and spectrogram visualizations. This practical assessment allows users to evaluate the model's performance in its intended use context and under realistic conditions. In the "My sounds" panel on this page, real-time predictions are displayed with a confidence score and bar chart below each class label (5.5b). Sounds recognized with confidence of 70% and above are added to a history panel on the right along with a timestamp (5.5c).[11] After testing, users may return to the previous

[11] The 70% threshold for sounds added to the history panel reflects a need for higher confidence in our study's 6-class models when compared to lower confidence thresholds used for larger-class models in prior work (*e.g.*, 50% [72], 60% [67]).

pages to add or remove sound labels, collect additional sampling data, or modify their training data and train a new model.

5.3 Evaluation

To evaluate SPECTRA and study how DHH users engage in IML to custom-train a sound recognition model, we recruited 12 DHH users for a two-hour, single-session remote user study. Participants used SPECTRA in their homes to create a personalized sound recognizer from their own soundscapes. Our primary research questions were:

- How do DHH users feel about sound recognition technology and the role of custom trained AI to improve and/or personalize sound classifications?
- How did DHH users engage with SPECTRA to interactively train a personalized sound model? Specifically, how did features like the waveform/spectrogram visualizations, interactive data clustering, and rich data annotations support and/or limit the IML process?
- How did using SPECTRA change perspectives about AI and their confidence in custom training a model?

Below, we describe our participants, the three-part study procedure, and our analysis method and positionality.

5.3.1 Participants

We recruited 12 DHH participants from two university-maintained study recruitment email lists and snowball sampling—see Table 5.1 for demographic details. Five participants identified as Deaf, five as hard of hearing, and two as deaf. Four participants reported using hearing aids; four used cochlear implants; two used both. We included two technology-related screening requirements: weekly laptop or desktop computer use and daily smartphone use for tasks other than phone calls and text messaging. Participants also needed access to a laptop or desktop computer at home with a working camera and microphone, a stable Internet

Table 5.1: Demographics of study participants. HH = hard of hearing.

ID	Gender	Age	Identity	ML Exp.	Hearing dev.	Relationship to sound
P1	Female	18 - 24	deaf		Cochlear imp.	*"Sound is extremely important to me in my daily life and in general."*
P2	Male	Over 55	Deaf		None	*"I was born profoundly deaf. I used hearing aids growing up but stopped since they didn't help me enough. I rarely relate to or experience sound."*
P3	Female	35 - 54	Deaf		None	*"I have to rely on sound awareness from my Hearing family members or friends to alert me. I'm very sensitive to loud vibrations, such as blasting from music, door slamming, or my iPhone beeping."*
P4	Male	25 - 34	HH		Both	*"I am hard of hearing with profound loss on one side so it is difficult to experience full sound unless I use my hearing aid"*
P5	Male	35 - 54	Deaf		Hearing aids	n/a
P6	Female	25 - 34	HH	yes	Both	*"I relate to and experience sound through hearing and hearing aids"*
P7	Female	25 - 34	HH	yes	Cochlear imp.	*"I'm no where [near] perfect hearing. [...] If someone is talking behind me I can only clearly hear a few words, not everything. If I am face to face, I understand perhaps 60-70% better than I did. Some folks can do better than me, and that's ok."*
P8	Male	25 - 34	Deaf	yes	Cochlear imp.	*"It's a love hate relationship. I love stuff like music and spoken languages enable me to relay technical information (where ASL has not caught up to yet) however I get listeners fatigue rather quickly [...] and have trouble distinguishing location without visual cues."*
P9	Male	18 - 24	HH		Hearing aids	*"I love music, but as my hearing has gradually declined, I've found it more difficult to notice and enjoy the musical aspects of life. When communicating with others verbally, I use context clues and patterns of speech intonation to help determine what words are possibly being said."*
P10	Female	25 - 34	HH	yes	Hearing aids	*"I generally experience sound as normal hearing people do except for speech. When I can't understand people it's not that I don't hear them, I just feel like they are mumbling and I can't understand them. "*
P11	Male	18 - 24	deaf	yes	Cochlear imp.	*"I use a Nucleus Cochlear Implant to hear sounds."*
P12	Male	25 - 34	Deaf		Hearing aids	*"I use sound to listen to music and to identify what is going wrong with my location."*

connection for videoconferencing, and Google Chrome to use SPECTRA. Nine participants used a laptop; three used a desktop. Participants received an $80 gift card as compensation for the 120-minute session.

We did not screen for prior ML experience among participants, as we designed SPECTRA to assist all DHH users interested in personalizing a sound recognition model. As a whole, participants self-reported moderate confidence in explaining basic principles of machine learning (*e.g.*, training data, predictions, models)—on a 7-point scale, the average response was 4.8 (*SD*=1.3, range=3-6). Five participants mentioned hands-on ML experience during the session (Table 5.1)—ranging from a brief university project (P6) to regular ML use for research (P8)—but none had worked on audio models before our study. Our study explored changes in their perceptions of audio models in particular (*e.g.*, tolerable performance levels before/after use), and we note users' experience where relevant in our findings.

5.3.2 Study Procedure (120 min)

The study session had three parts: (1) introducing participants to technical concepts and SPECTRA; (2) using SPECTRA to record data, train, and test a personalized sound recognition model; and (3) a semi-structured interview discussing the experience. Before the study, we administered an online questionnaire to collect demographics, use of sound support technologies, confidence in explaining ML concepts, and prior experience with smartphone sound recognition tools. The first author led all interviews remotely using videoconferencing software with automatic captioning enabled. Participants could request sign language interpreting or real-time captioning accommodations; four opted for interpreters to join the call. Participants received consent forms via email and provided verbal consent at the start of the session.

Tutorial and Pre-Use Interview (30 min)

Sessions began with five minutes for Zoom setup and orientation, followed by a tutorial slide deck[12], which participants could navigate at their own pace. The tutorial, informed by prior work on IML systems for non-expert users [122, 148], provided an overview of sound recognition models' learning and decision-making, the differences between generalized and personalized models, and possible advantages of personalization. It then introduced each stage of the SPECTRA pipeline, with accompanying screenshots demonstrating recording, training data selection, model training, and assessment as well as explanations of spectrogram, waveform, and data clustering visualizations (*e.g.*, *"Examples closer together are more similar, while those further apart are more different"*). We encouraged participants to ask questions throughout the tutorial.

Upon completing the tutorial, participants responded to rating scales measuring their self-reported confidence in recording, training data selection, assessment, and using SPECTRA, along with their performance expectations and error tolerance. Each rating consisted of a subjective statement (*e.g.*, *"It's okay if a sound recognition model that I have trained occasionally makes mistakes"*) followed by a 7-point agreement scale from *"Completely disagree"* to *"Completely agree"*. We then conducted a brief interview to elicit feedback on the benefits, drawbacks, and desired sounds for a personalized sound recognition model, along with strategies for capturing diverse examples and their tolerance for model errors.

[12] The full tutorial slide deck is available in Supplementary Materials

System Use (60 min)

After completing the tutorial, participants accessed SPECTRA via a shared link and began screen-sharing with videoconferencing. We instructed participants to “think aloud” throughout system use and to freely voice any questions, observations, suggestions, or concerns that arose. For participants using ASL, comments followed or preceded system interactions rather than occurring concurrently. To prevent significant barriers to training a model, the researcher provided troubleshooting and clarification support where appropriate; we noted these areas of friction and included them in our analysis.

We asked participants to train a model with five sound classes, plus “background noise” to provide a baseline for the ambient soundscape—a decision informed by Jain *et al.*’s survey in which a majority of Android sound recognition users desired notifications for six sounds or fewer in a single location [67]. To orient participants to background noise in their environment, we asked them to turn on their microphone and observe the visualizations, along with their changes when intentionally making noise. We defined the soundscape in the absence of intentional noise as “background noise” for the purposes of this study. We pre-selected three sounds (door closing, faucet, and phone) based on how easily they could be produced and the range of possible variations. Participants brainstormed and then selected the remaining two sounds.

Data collection. Participants began by collecting at least 30 seconds of audio for each sound. The researcher guided participants through recording background noise, and participants independently collected the remaining sounds. We encouraged recording from the environment where possible, using the sound library only if needed; no participant ultimately chose to collect audio from the sound library. We instructed participants to aim to “capture the real-world variations” that may occur for each sound and prompted them to consider and record how sounds could happen differently in their home (e.g., “running your faucet on full vs. a light stream”), We reminded them to use annotations to track any recorded sound variations (*e.g.*, faucets in different rooms) or other relevant details. Because prior work [71], found that distance from the microphone to the sound source does not significantly impact model performance, we did not emphasize recording location as a key variable to consider. However, some participants moved their devices throughout the home while recording. After collecting data, we offered a 5-minute break before continuing.

Model training. Participants then moved to the “Train” tab to filter their sampling dataset into a training set. We instructed participants to generate an initial clustering for the full dataset first and share their observations of outliers, overlapping sounds, or well-separated classes they observed. We then asked them to review each sound class, removing any examples they believed were unsuitable for the training set while explaining their reasoning. The researcher periodically prompted participants to generate a new clustering chart after making changes to the selected dataset then discussed any perceived changes and perceptions of its implications for their model. We encouraged participants to continue refining their training data set until they felt satisfied, probing participants about what they were observing that drove decisions to remove or include data. Participants leveraged both the visualization of each sample and the clustering visualization to reason about in/exclusion. Once satisfied with their training set, participants trained a new model with this data.

Model testing. Proceeding to the final tab (“Test”), the researcher asked participants to assess their model’s real-world capability for everyday use by reproducing each sound for at least 10 seconds, reminding them of any variations they had identified earlier. Participants discussed the model’s output, theorizing about potential reasons for misclassifications and possible fixes to improve performance. After testing each sound, participants could use their remaining time to return to the previous tabs to adjust their model as desired (e.g., record more data, continue refining the dataset)

Semi-Structured Interview and Rating Scales (30 min)

We concluded the study with a post-use questionnaire and semi-structured interview. The questionnaire measured changes in confidence and performance expectations after use (mirroring statements on the pre-use ratings); satisfaction with recordings, training data, and model accuracy; and the usefulness of the text annotations, waveform, spectrogram, and clustering visualization. The interview explored overall satisfaction, experience with each step of the application, confidence in independent training, and opinions on the data exploration mechanisms and potential UI improvements.

5.3.3 Analysis and Positionality

We collected each participant's usage logs, audio recordings, and training datasets to further characterize their responses and experience with the pipeline. We used Zoom's automatic captioning to transcribe study data, relying on voiced interpretations as an accurate representation of signers' responses. We iteratively coded transcripts using reflexive thematic analysis [14, 15]. Our analysis was semantic and realist, and we developed themes using a mixed inductive and deductive approach; for example, we structured broader theme development around the distinct tasks at each step of personalizing a sound recognizer (*i.e.*, recording, choosing training data, testing), but we organically identified themes within each step. The first author read through the data, generated initial codes, and then applied these codes to data from two randomly selected participants. Another researcher reviewed the coded data and then met with the first author to discuss the codes and application strategy. The first author coded the remaining transcripts and generated themes from data excerpts collated from each code. A reflexive approach to thematic analysis emphasizes findings that are actively shaped by the research team's own social, cultural, and academic biases. All authors are hearing, while past collaborators—who contributed to the early system and study design—are Deaf or hard of hearing. All research team members have backgrounds in human-computer interaction, and many are computer scientists by training.

5.4 Findings

We present findings organized around our primary research questions: (1) pre-use expectations and feelings about sound recognition technology; (2) engagement and use of SPECTRA to interactively train a sound recognition model; and (3) post-hoc reactions to the experience, including self-confidence, performance tolerances, and technical understanding. We begin with an overview of SPECTRA's usage, the collected data, and model training.

Table 5.2: Collected data and system usage for study participants. (Column A) Participants chose sounds to train in addition to the required "background noise", "door closing", "faucet", and "phone" sounds. (B) They captured at least one n-second recording of each sound. (C) Their recordings were segmented into 1-second examples to use as training data. (D) They selected a subset of the examples to train a model. (E) They generated clusterings to visualize different iterations of this subset.

ID	**(A) Sound choices**	**(B) Total recordings**	**(C) Total examples**	**(D) Training examples**	**(E) Clusterings generated**
P1	Microwave running, Knocking	10	205	191	2
P2	Door knock, Washer/dryer signal	10	265	215	6
P3	Dryer, Stove Exhaust Fan	6	174	160	2
P4	Fridge, Vacuum, Printer	12	280	255	4
P5	Keyboard, Door knock	6	214	172	4
P6	Zipper, Male voice	12	218	202	6
P7	Door knock, Keyboard typing	8	250	238	4
P8	Microwave, Footsteps	16	263	212	4
P9	Knocking, Garage door	9	422	257	6
P10	Stovetop fan, Blinds	13	327	212	5
P11	Vacuum, TV	11	284	261	8
P12	Door knocking, Shower	7	341	254	3

5.4.1 Overview of Collected Data and SPECTRA Usage

All participants were able to train a personalized model using SPECTRA. In total, participants captured 120 recordings across 70 sound classes[13]—see Table 5.2. The most common created sound class was "Door knock"/"Knocking" (*N*=6/12) followed by "Microwave", "Stove fan", and "Keyboard" (*N*=2 each). Participants collected an average of 1.7 recordings per sound, with an average duration of 27.0 seconds (*SD*=4.6, range=2-104). Recordings were automatically segmented into 1-second Mel spectrograms, resulting in 46.3 examples per sound on average (*SD*=10.3, range=30-104). To improve model robustness, we instructed participants to consider different ways sounds could happen in their homes. A few participants captured this variation within a single recording (*e.g.*, P12's annotation: "Shower with many varieties"), but most chose to collect separate, annotated recordings (*e.g.*, P10: "faucet low"/"medium"/"high stream").

For training and testing, participants spent an average of 20.7 mins (*SD*=5.7) on the "Train" page and 10.6 (*SD*=2.0) on the "Test" page. The training itself was interactive and iterative, with nearly half of the time spent focused within the clustering chart—on average, participants clicked on 7.5 clustering

[13] Two participants (P3, P5) removed "Phone" due to issues producing the sound (*i.e.*, an alarm or ringtone). P9, a desktop user, replaced "Faucet" with "Printer" due to proximity.

points and regenerated clusterings 4.5 times (*SD*=1.8, range=2-8) after making changes to their training dataset. Participants removed an average of 8.6 examples in their final training datasets (23% reduction), demonstrating the visualizations' influence on their decision-making. Overall, final training datasets averaged 37.7 examples per sound (slightly above the minimum requirement).

5.4.2 Pre-Use Perceptions and Expectations

In the pre-study questionnaire, most participants (*N*=7) expressed positive interest in automatic sound recognition technology ("likely" or "extremely likely" to use it), including for urgent (P7: *"smoke alarm"*), social (P1: *"someone arriving home"*), and appliance sounds (P9: *"oven timers"*)—aligning with prior work [13, 37]. Four participants remained "neutral", while P6 was "unlikely" to use such technology. Participants also identified several uses for personalized models, most commonly related to identifying specific speakers (*N*=5) and nuanced pet sounds (*N*=4).

Five participants reported using sound notification features on their smartphones, albeit infrequently (semi-monthly or less), citing limited sound selection support and inaccurate recognition as key issues—echoing past findings [67]. Only one of the five had previously attempted to add a custom sound class (on the iPhone), but even here, the participant experienced issues: *"It's like, 'Someone's knocking at the door,' but it's actually my roommate, cutting with a knife"* (P5).

Upon completing the tutorial, participants expressed confidence that they would be able to create a sound recognition model (*avg*=5.9, *SD*=1.0): *"I feel pretty good—the tutorial and the way you made [SPECTRA] makes it seem pretty straightforward"* (P10). They also had moderate expectations that a personalized model would identify sounds accurately (*avg*=5.0, *SD*=1.3) and emphasized the tool's value, even with mistakes: *"[It] might [still] be a significant benefit over what my baseline is"* (P8). However, this optimism was tempered by uncertainty due to their unfamiliarity with SPECTRA, machine learning, and/or the effort required to complete the workflow: *"30 seconds per sound; it sounds like a lot of work"* (P4).

5.4.3 Using SPECTRA to Interactively Train a Personalized Sound Recognition Model

We describe how participants used SPECTRA's waveform and spectrogram visualizations, interactive data clustering, and rich data annotations to train a personalized sound model.

Waveform and Spectrogram Visualizations

Sound visualizations are essential for making audio data accessible to DHH users [47, 107]. However, how best to visualize sound to support interactive training of a sound recognition model is an open research question—especially for users who may have different mental models of sound and/or lack access to the auditory channel itself. Thus, drawing on prior work [47], we designed SPECTRA to use waveform and spectrogram visualizations for both streaming and static sound information when recording audio, selecting training examples, and testing a new sound recognition model.

In general, waveforms were rated as highly useful for recording and reviewing examples during the IML process (*avg.*=6.7, *SD*=1.2). Participants found the waveform to be intuitive (P3: *"Like one of those heartbeats on the EKGs"*), clearly showing that sounds were captured (*e.g.*, P12: *"I can see the microwave, [...] the four beeps"*) or if unwanted sounds occurred, such as *"my cough"* (P6). The waveform's shape and amplitude helped participants build intuition about the model's classes, highlighting distinctive characteristics of sounds (*e.g.*, short "Door closing" *vs.* sustained "Faucet") and the effects of controllable variables like speed, intensity, and distance (*e.g.*, P6: *"I can see the difference when I closed the door very hard, it's more thick"*). When constructing a training dataset, the waveform's glanceability proved especially useful for scanning the selection grid on the "Train" page to identify examples for removal; P4 noted, *"The background noise [vs.] whenever I was talking, being able to figure out which one was which—I think that was really helpful."*

While less preferred overall, participants also found the spectrogram useful for reviewing collected audio (*avg.*=5.1, *SD*=1.6). *"The spectrogram, it's useful too, but it's not more important than the waveform"* (P6). Most felt the spectrogram was less intuitive than the waveform—*"I don't identify things in my life really based on frequency as much as I do based on loudness"* (P9)—and some even found it *"confusing"* (P1). However, the spectrogram proved useful to a few participants for in-depth reasoning, such as P11: *"My concern is with the [ringtone], it looks too similar to the sink faucet. This is probably getting mixed up"*. Just one participant, P10, preferred spectrograms to waveforms; she used it to reason about inaccurate predictions: *"It said [the faucet sound] was maybe blinds. The blinds [spectrogram] had a lot of bands which were more high frequency. [...] The slower [faucet] drip, I think, looked similar to that"* (P10).

Interactive Data Clustering

Participants deemed the data clustering visualization critical to the IML process—usefulness rating: *avg.*=6.3, *SD*=1.2—primarily because it helped provide transparency, feedback on audio recording quality, and assisted with iteratively refining the training dataset. As P7 stated, *"[It's] a good depiction of where everything lies and how the model is looking at it"* (P7). Participants saw clusters as useful when trying to understand the consistency of samples in a given class as well as distinctiveness across classes: *"This is the door closing, and it's clustered right here, so I know that I did a good job"* (P8). Clustering also served as a bridge for participants to begin considering their data in terms of how it may impact the behavior of a ML model – P2 iterated on their data set until the clustering seemed *'more clear [now]; it doesn't seem as if there's a lot of overlap"*. For P1, watching the clustering change after adjusting the training dataset felt affirming: *"[It] was satisfying to see, 'Okay, like it's actually working; what I'm doing'"*). In contrast to prior work where DHH participants expressed uncertainty about the quality of self-collected training data [47], using the clustering visualization increased participants' confidence that they had collected their desired data; upon a final review of his clusterings before training, P4 said, *"I feel pretty good about this—the [examples] that are remaining."*

One participant, P9 (HH, hearing aids), felt that the data clustering visualization was *not* useful after struggling to clearly separate his sound classes. Two key problems emerged: first, he captured his new class—a "Garage Door" sound—at a distance, which created a noisy sample. Though P9 directly observed this issue in the recording visualizations (*"It feels like it's getting something, but it's really tiny [in the visualization]"*), he initially did not understand its impact on class separation: *"Even when I removed sounds that based on the waveforms and spectrograms don't seem to matter, [the sounds] are still really bunched together. [...] 'How can I fix this? Do I need a new [garage] door?' Well, these sounds aren't gonna change."* Interestingly, though frustrated, P9 eventually diagnosed the problem: *"I might just give up on sounds like the garage door, just because they're too close to background noise and it didn't differentiate it. I need to add some* *more distinct sounds."* P9's struggles highlight both the importance of good training data and an opportunity for clustering visualizations to assist users in learning to reason through the differences between human and machine perception of audio.

Rich Data Annotations

All participants agreed that data annotations were useful (*avg.*=7, *SD*=0). Because recordings were auto-labeled with their sound class, participants did not need to add specific data annotations. But all did, and over 83% (*N*=100) of recordings included an annotation. Most annotations (*N*=73) emphasized differences in the sound's production (*e.g.*, P1: "quiet knocks" / "louder knocks") or the recording's proximity/location (*e.g.*, P11, vacuum: "far" / "near"). Several annotations (14) instead noted contextual information about the recording, such as the presence of co-occurring sounds (*e.g.*, P6: "close door so hard and my husband's voice appeared") or other helpful context (*e.g.*, P5: "the faucet started in the middle"). The remaining annotations (13) were procedural (*e.g.*, P2: "Phone" / "Phone #2").

Training strategies.

We identified two strategies used by participants to incorporate the clustering feedback into their training data choices. In the first strategy, participants saw the clustering visualization as a method of checking progress while filtering out unhelpful training data, but relied on spectrogram or waveform visualizations of each sample to judge the quality of training examples. With this approach, participants flagged individual examples for removal by comparing their visual shape to the other examples of that sound. Often, this was as simple as noting flat waveforms (*e.g.*, P2: *"It's important to remove the lines that are quiet"*), but sometimes, it involved a nuanced judgment of the visualization's meaning by recalling the recording's context. For example, after P10 noticed *"something was different when it started"* for a faucet recording, she reasoned, *"It's probably the water just hitting the sink,"* and ultimately chose to include the dissimilar example in her training set. After removing one or several examples from the training dataset, they generated a new clustering chart to see how their overall training dataset had changed (*e.g.*, P5: *"It's still a little mixed, but it does seem like the [background noise] is now pulled apart a little bit, and the [faucet]"*).

In contrast, the second strategy leveraged the data clustering chart as an interactive tool to identify problematic examples. The participants who used this strategy primarily searched for individual examples *"outside of the group"* (P12), embedded far from the other examples sharing its label (outliers). Upon selecting an outlying example, they turned to the visualization and reasoned about its contents (*e.g.*, P4, background noise: *"I was maybe talking"*; P6, phone: *"I put [my phone] on the table"*) to decide whether or not to exclude it from the training dataset. Participants taking this approach said the clustering visualization *"helped me to make more sense of the data, but I think more so, it helped to guide me in [the] refinement process"* (P8). P11 further explained an efficiency benefit: *"I was driven by what I was seeing in the chart [...] to eliminate some edge cases and anomalies. [...] Everything is [shown] together, but in [the selection panel], I have to compare one by one"*. After removing an example, they updated the data clustering chart and searched for new outliers, repeating this cycle until none remained. Here, participants believed that samples that appeared as outliers in the clustering visualization represented samples that would not result in a high-quality model.

Design Suggestions

Participants shared suggestions to improve SPECTRA and IML workflow. Some participants (P2, P6) suggested that it would be useful to record sound on a smartphone but continue interactive model training on a laptop/desktop, to balance mobile capacity with a the affordances of a larger screen. Participants had suggestions to simplify the workflow, as they felt that SPECTRA required too many interactions (*e.g.,* unchecking samples) to produce a useful result. Others felt that creating an entirely new model was unnecessary, preferring to *"append new sounds"* (P11) to an existing model instead. Finally, participants desired more direction mid-use. For *in situ* help, participants suggested *"text reminders"* pointing out problematic examples (P7), *"tips about what to look for"* (P10) in the clustering visualization, and a persistent *"guiding hand"* (P8) to offer suggestions and assistance throughout the model-building process.

5.4.4 Post-Study Questionnaire and Interview

Following the IML task with SPECTRA, we concluded the study with a questionnaire and semi-structured interview. We describe participant reflections on using SPECTRA, including reactions to model performance, handling and understanding errors, training strategies, and new suggested use cases.

Overall Perceived Usefulness.

Overall, participants felt that personalized sound recognition and model training was useful and *"applicable to daily life"* (P1), voicing intent to *"look into using [it] if it becomes widely available"* (P10). They noted new possibilities for personalized sound recognition models, including *"auditory pedestrian traffic signals"* (P7), *"a car alarm"* (P10), and *"[my] baby crying"* (P11)—while P2 said, *"I would want to record everything"*. P1 described this newfound sense of agency: *"It's just kind of exciting [...] that it can recognize these specific sounds and be trained, and it's accessible to people like me."* However, for some, like P8, a personalized model did not seem to provide advantages over his existing sound awareness adaptations: *"I have residual hearing, I use a cochlear [implant], so I can probably hear these anyways. [...] doors closing and footsteps, I'm going to feel the vibrations in the house."*

Task Approachability and Self-Confidence.

Though most participants lacked experience with machine learning, by the end of the study, all felt confident in personalizing a sound recognition model with SPECTRA (*avg.*=6.3, *SD*=0.9). As P10 expressed, *"I was kind of surprised that it actually worked—it's just cool to see"* (P10). Most found the workflow well designed; *"It [was] relatively self-explanatory once you fiddled with it"* (P9) and P3, who was originally timid, *"loved it at the end"*. Participants felt most confident about data collection (*avg.*=6.3, *SD*=1.2) followed by model training (*avg.*=6.2, *SD*=0.9) and testing (*avg.*=6.2, *SD*=0.6). Data collection was *"pretty simple"* (P3), *"convenient"* (P9), and *"unique"* because *"I don't normally think about [these sounds] in terms of recording"* (P7). However, the training and testing stages were harder: P4 indicated *"not understanding"* at first. Most cited the visualizations as useful and learned to judge data quality themselves; *e.g.*, *"[The clustering supported] understanding of what's happening under the hood"* (P8). Though the "Testing" tab was *"well-designed"* and *"comfortable"* (P11), some participants struggled to understand how to improve model performance.

Reactions to Perceived Model Performance.

In general, participants felt that their personalized models classified sounds accurately (*avg*=5.3, *SD*=1.4). In several cases, model performance exceeded participant expectations. For example, although P2 initially wanted to *"skip [testing] 'Door knocking'"*, assuming it was*"just going to overlap"* with 'Door Closing', his custom-trained model successfully discriminated between the two sounds: *"Awesome, I think it was accurate."*. And P4 was *"very satisfied"* with his model, despite having low expectations due to prior experience with Android's *Live Transcribe*: *"I'm very satisfied with how it turned out, but I think if I hadn't been exposed to [Live Transcribe], then I would have a higher bar."* However, P9's high initial expectations were tempered after *"learning more about the process"* and understanding that *"it can't pick up all [the] sounds."*

Handling and Understanding Errors.

In the pre-use questionnaire, participants acknowledged that some sound recognition errors are likely unavoidable, and these perspectives remained consistent throughout the session. A few participants mentioned false positives as more tolerable than false negatives before using SPECTRA. P11 maintained this perspective while testing his model, even after it mistook his phone vibration for "TV" and "Door closing" sounds: *"I'd be okay with [that] because it tells me something's happening around the house."* P9 *"really liked"* confidence scores displayed with the model's predictions, as it allowed him to reason about misclassifications using his residual hearing abilities: *"If it's at 100%, maybe I heard [the sound], but if it's at 75%, maybe I didn't, so maybe I should look more into it. And if it didn't come up [on the screen] and I feel like I heard it, then it's not [working]."*.

Strategies to Improve Model Performance.

Using SPECTRA shaped participants' perceptions of and intentions for future use of IML for personalized sound recognition. After using SPECTRA, seven participants' self-reported understanding of how to improve model performance actually declined (five ratings increased). Having grappled with the complexity of training a sound recognition model, participants had ideas for what they would do in the future. Note that while iterative model refinement is a cornerstone of IML pipelines [30], none of our participants trained a second model due to time constraints or fatigue. Some participants imagined how they would arrive at more practical sound classes, either adding *"more [...] things that are important to me"* (P7) or removing less valuable sounds, such as *"the door opening and closing, I don't think I need that"* (P2). Non-expert users commonly believe they will see performance gains by adding more data [148]; a few participants voiced similar ideas. Beyond the kind of data collected, participants saw promise in data refinement, such as removing outliers in the clustering chart (P10: *"the little points there, sticking out"*) or recording additional sound variations to introduce edge cases; for example, P1 was *"curious"* how her model would respond to *"different kinds of doors or different ringtones"*. Upon reflection, participants had new ideas for sound classification schemes, informed by their experience observing overlaps and differences in the classes they tested. After misclassifications during testing, P12 sought to resolve *"overlapping sounds"* in his clustering chart, realizing that: *"Shower and the faucet [...] maybe I could combine and have 'water running'"* (P12). P4 sought to improve the performance of his "Fridge" class; recalling its two separated clusters, *"[I would] focus on just the 'ice' [dispenser], just because the 'water' [dispenser] is similar enough to the 'Faucet'. [...] [I'm] confusing the model by having two different sounds come out of the same object."*

5.5 Discussion

Our work advances understanding of how to support DHH users in training personalized sound recognizers by: (1) investigating non-auditory data representations across an end-to-end training cycle for data collection, training data selection, and practical testing; (2) demonstrating how interactive data clustering can support DHH users to reason about audio data, identify outliers, and refine training datasets; and (3) provide insights into DHH users' experiences and perspectives on personalizing a sound recognition pipeline.Our work also reaffirms prior work showing DHH users' preference for waveforms when recording [47]—while expanding on their value when selecting training data and testing—and how training strategies can change through use [107]. Below, we situate our findings in the literature, offer design considerations, and discuss limitations and opportunities for future work.

5.5.1 Design Considerations for Interactive Sound Recognition Tools

Based on our evaluation of SPECTRA, we share the following design considerations for future tools:

Interactive clustering visualizations. We found clustering effectively provides non-auditory feedback on interclass relationships, supporting DHH users' understanding of an audio dataset (a key challenge identified in prior work [47]) as well as their iterative refinement of a training subset. In this way, the visualization enables more active participation in model-training—another challenge for DHH users [107].

To better highlight the impact of training data inclusion or exclusion, participants requested visual cues or side-by-side comparisons. Designers should also be mindful of potential overfitting when users rely solely on clustering for training data selection. Future work could investigate how inclusion and exclusion decisions may impact model performance and provide user feedback accordingly. To further mitigate overfitting, encourage users to focus on outliers and overlap and emphasize clustering as a representation of the model's perspective rather than the ground truth of decision boundaries.

Waveforms. Our findings suggest that waveforms are essential for DHH users throughout the IML workflow, and their single dimension of amplitude (loudness) *vs.* time is intuitive for this population. For DHH users to monitor sound input and their soundscape, waveforms should be displayed prominently before and during recording (extending prior work [47]) and when testing models. When selecting training data, waveforms offer a glanceable representation that provides transparency into individual audio examples and adds context for locations within the clustering visualization.

Spectrograms. While spectral features are the standard input for sound recognition models [55], spectrogram visualizations did not offer a significant benefit to DHH users in our study. In contrast to the waveforms' simple vertical amplitude, spectral information depicted by frequency on the vertical and amplitude as color intensity is less intuitive—even confusing—for DHH users. However, spectrograms may offer limited value for in-depth analysis when selecting training data (particularly to experienced users [134]). Other time-frequency visualizations, such as correlograms or pitchograms [20], may offer more value as visual analysis tools for this population and present an opportunity for future work.

Annotating. Our findings suggest that allowing DHH users to provide notes about a sounds' production, location, and context aids their understanding and ability to use an IML workflow. Some annotations drove users' exploration of nuanced subcategories within a sound class; designers can proactively support this by suggesting potential subcategories from the start (*e.g.*, generating subcategories for a given sound class via a language model). Highlighting annotated subcategories visually in the clustering (*e.g.*, P11: two discrete clusters for different phone ringtones) can further expose distinctions in the model's decision space, aiding comprehension. Concept decomposition options [117] can streamline clustering insights—either to separate disparate clusters into their subcategories (*e.g.*, P4: fridge → water, ice dispenser) or to combine overlapping classes (*e.g.*, P12: faucet, shower → water running).

Multiple views of information. Our study highlighted DHH users benefited from the interplay of multiple views of sound data: clustering provided high-level structure, waveforms showed individual example content, spectrograms offered nuanced detail (to some), and annotations supplied context. Future systems could incorporate multimodal information, such as allowing users to capture video recordings of sounds for an additional analysis dimension that leverages users' visual reasoning and memory.

System format. Participants wanted to capture data on mobile devices but requested flexibility in the device for IML; cloud-based applications can allow users to take a multi-device approach. When adapting IML workflows to mobile formats—the device form factor that is ultimately preferred for daily sound awareness [37]—prioritize waveforms for data collection, clustering throughout training, and waveforms + predictions during testing.

5.5.2 Future Opportunities for Efficiency and Model Optimization

We found clear limits to the time and effort that DHH users are willing to invest in personalizing sound recognition models, creating a tension with streamlining personalization tasks for efficiency without reducing users' engagement in an interactive training process [3, 30]. Prior work [67] found that Android sound recognition users hoped to spend less than 25 minutes on personalization; training a model with SPECTRA required considerably more time—the allotted hour for most users—and those with time remaining declined to retrain their model. While users found the clustering visualization highly engaging, data selection was a key area to streamline: improved data processing (*e.g.*, automatic segmentation, silence removal) can reduce data cleaning efforts, while automatic outlier detection (*e.g.*, [121]) can highlight atypical examples for review.

Optimized ML architectures or extensions to pre-trained models can further reduce the effort required for interactive personalization. While SPECTRA's Speech Commands API [136] was well-suited for prototyping interactive training workflows for our study, Jain *et al.*'s ProtoSound is a more optimal architecture for the daily needs of DHH users (*e.g.*, contextual flexibility, open set classification) [67]. Protosound combines few-shot learning with prototypical networks to train custom sound models that outperform comparable architectures (including with DHH users' recordings); however, ProtoSound's "black box" interface lacks audio visualizations and control of the training dataset. As SPECTRA focuses specifically on frontend support, combining it with ProtoSound is a clear next step to improve the baseline performance of users' models, reducing time for model refinement. For example, SPECTRA's clustering visualization integrated within the ProtoSound pipeline could include embeddings generated from the model's internal feature representations, yielding true insight into how the model differentiates sounds and its decision-making process. Finally, our findings also highlight that rather than create a new model, some users feel that adding custom classes to existing sound models is a simpler task. Similar customization features are supported in iOS (tuning existing classes) [7] and Android (adding new classes) [51], but these, too, lack accessible data representations and training insight; supporting the interactive extension of pre-trained models is an opportunity for future work.

5.5.3 Limitations

Our work has several limitations. First, we did not conduct formal analyses of the participants' models and thus cannot definitively quantify the impact of SPECTRA's accessibility features and participants' decisions on model performance. Further, the quality assessment stage was limited to reproducing their model's trained sounds within the system and did not include metrics about the model; as a result, participants' high opinion of their models may have been inflated due to the lack of long-term practical use.

Second, while our evaluation presented SPECTRA within a familiar, domestic soundscape, we acknowledge that new challenges with user-driven personalization may emerge in complex acoustic environments with many similar and/or overlapping sounds. Future longitudinal studies are needed to investigate interactive training and deployment within busy or unfamiliar locations—settings where DHH users may be in the most need of sound awareness support [37, 47].

Next, while we did not aim to recruit people with ML expertise, five of the 12 participants had hands-on ML experience. Though often not extensive and not in the audio domain, this experience suggests that our participants may not represent the general public, limiting the generalizability of our findings.

Finally, we had limited control of the testing environment: our remote evaluation using videoconferencing software led to greater setup and troubleshooting time while reducing opportunities for retraining, experimentation, and/or discussion for some participants. The abbreviated experience may have impacted participants' opinions about their models, and future longitudinal studies can better explore how users' perceptions change through continued use.

5.6 Chapter Summary

This chapter explored the design and evaluation SPECTRA, a prototype for the accessible creation of personalized sound recognition models, merging interactive ML guidelines with the needs of DHH users across an interactive training workflow. We evaluated the prototype in a hands-on model-training session with 12 DHH participants; our findings highlight the potential of interactive clustering visualizations to support DHH users in exploring the composition of personal audio datasets (a key challenge identified in the previous chapter). Further, we detailed how users combined multiple information streams (clustering, waveforms, spectrograms, and annotations) to reason about the suitability of their training data and meaningfully engage in the training process. Our observations of participants' training successes and challenges offer valuable insights that, in combination with those from the previous chapters, form an empirical understanding of DHH users' needs and preferences across key steps of an interactive sound recognition pipeline (problem-framing and feedback from trained systems, data collection, and model training).

Chapter 6

Conclusion & Future Work

This dissertation constitutes a framework for supporting DHH individuals to personalize sound recognition tools that meet their everyday needs. It accomplishes this in two ways: (1) building an empirical understanding of DHH users' needs and preferences within key areas of an interactive sound recognition pipeline (problem-framing and feedback trained systems, data collection, and model training), and (2) guidelines for the design of future personalizable sound recognition systems to meet this understanding. In this chapter, I review the empirical and design contributions of my dissertation, discuss possible directions for future work based on these efforts, and outline its limitations.

6.1 Empirical Contributions

My dissertation makes several empirical contributions to the fields of human-computer interaction and accessibility, specifically toward understanding DHH individuals' needs and preferences around personalization in sound awareness tools.

First, I investigated the utility of different forms of sound feedback for DHH users and how contextual factors can influence the relevance and perceived value of that feedback (Chapter 3). The findings showed that DHH users generally prefer multimodal feedback in a wearable sound awareness system: visual feedback offers detailed information and glanceability, while haptic feedback provides discreet and immediate alerting—a necessity for DHH users who otherwise rely on visual awareness. An *in situ* exploration showed that DHH users prefer to filter sound notifications in busy environments, but no single filtering method (by identity, loudness, or direction) was preferred for all situations. Additionally, some were uncertain about sounds being filtered automatically—indicating a need for personalizable feedback and filtering options to cater to individuals' sound interests.

Second, I explored the practical considerations and sense-making strategies that DHH people use in recording and interpreting real-world audio data for training a sound recognition model (Ch. 4). When choosing sound classes, users considered possible decision boundaries and appropriate diversity within their samples, but inexperience with ML and each sound's real-world distribution led to decisions based on guesswork. Continuous, prominent, and controllable sounds were easy to sample, but spontaneous, invisible, and complex-to-produce sounds were more difficult—even impossible—for users to capture. Key strategies to interpret the samples' contents included the real-time waveform, listening to audio playback (for those with residual hearing), and comparing shapes of post hoc waveform visualizations; however, users' limited experience with the sounds led to breakdowns in the waveforms' meaning and uncertainty over the representativeness of their samples. They requested features that could better inform about soundscape activity—especially for co-occurring sounds—and insight toward how a model might interpret each sample in relation to their larger training set.

Third, I evaluated SPECTRA, an end-to-end prototype for interactive and accessible model-training, yielding insights on DHH users' training strategies and conceptualization of ML when creating personalized sound recognition models (Ch. 5). My findings demonstrated that interactive clustering visualizations can help DHH users to better understand the structure of their personal audio data and make decisions to refine their training datasets. Participants combined information from clustering, waveforms, spectrograms, and data annotations to reason about their audio data's suitability for training their models, leading to a nuanced understanding of the relationship between training data and observed model performance. The interactive workflow led to more active and informed participation in the model-training process, which is critical for interactive ML systems but has been a challenge for DHH users in prior work. However, the evaluation also revealed a tension between training time and effort and user engagement in the personalization process; striking a balance in these areas as a key concern for future work.

6.2 Design Guidelines for Future Personalizable Sound Recognition Tools

In addition to empirical contributions, each chapter of my dissertation contributes guidance for designing personalizable sound recognition technology. In particular, these guidelines adhere to DHH users' preferences for sound feedback and filtering, their experiences recording personal audio data, and their needs when interpreting and assessing this data to train personalized sound models.

Sound Feedback & Filtering. My contextual exploration in Chapter 3 revealed complementary roles for visual and haptic sound feedback, along with the necessity of filtering to manage complex soundscapes.

1. **Visual and haptic feedback should be used together**, with simple visual designs on the smartwatch for glanceability and haptic feedback for attention-getting without interrupting visual awareness.

2. **Tactons (haptic patterns) can be used to convey sound information**, but they should be limited to a small, configurable set to manage the learning curve and user preferences.

3. **Filtering options should be provided to manage soundscape complexity**, allowing users to filter by sound identity, loudness, and/or direction, and to switch between filtering presets based on their context.

4. **Systems should be transparent in filtering and identification decisions**, providing real-time information and allowing user modification to foster trust and address potential accuracy concerns.

Recording Support. My field study in Chapter 4 provided implications and considerations for designing specialized recording tools to aid DHH users in capturing a personal audio dataset.

1. **Include combined scaffolding for ML and audio concepts** that considers the limited auditory experience of DHH users, such as explanations of sound features (*i.e.*, spectral data), examples

of sounds that may have overlapping decision boundaries, and a breakdown of a sound model's decision-making processes.

2. **Sound visualizations, such as waveforms, are crucial** to reveal soundscapes and the contents of recordings; however, multiple visualization options are needed (*e.g.*, spectrograms) to minimize breakdowns in users' intuition, support comparison strategies, and help uncover deeper insights about recorded audio.

3. **Provide high-level feedback for auditory information** that DHH users might be less experienced with or unaware of, such as the diversity of their dataset, detection of co-occurring sounds, or other information about the ambient soundscape.

4. **Allow for multiple data sources beyond user-captured examples**, such as sound libraries, to account for sounds that are difficult or impossible for DHH users to capture themselves.

Training Workflow. In Chapter 5, I built an end-to-end prototype to assist DHH users with training a sound recognition model, leading to recommendations for UI elements that can support DHH users in interpreting their audio data when training and evaluating a sound recognition model.

1. **Employ interactive clustering visualizations** to help users identify outliers, understand the impact of their training data selections, and monitor their progress in optimizing the model's performance. To mitigate potential overfitting, emphasize clustering as a representation of the model's perspective rather than the ground truth of its decision boundaries.

2. **Allow users to annotate their data with contextual information**, supporting their understanding and exploration of nuanced subcategories within sound classes, with straightforward options to merge or separate classes during model refinement.

3. **Support multimodal data exploration** by allowing users to leverage different data representations (*e.g.*, waveforms, spectrograms, annotations, and even video recordings) to build a comprehensive understanding of their dataset and carry out informed model-training.

4. **Minimize the time and effort required for data collection, cleaning, and model training** by employing techniques like automatic segmentation, outlier detection, and optimized machine learning architectures. Systems should seek to balance efficiency with user engagement.
5. **Offer a multi-device system format** for users to capture audio on mobile devices, with the flexibility to utilize other devices for model training, evaluation, and deployment.

6.3 Opportunities for Future Work

My dissertation explored the immediate stages of the machine learning workflow (feedback preferences, data collection, training, and testing), but the success of personalizable sound recognition technology lies in its long-term adoption and use. To this end, future research should focus on the following areas:

Model Deployment and Ongoing Refinement. *How do DHH users integrate personalizable sound recognition tools into their daily lives, and how do their perceptions and attitudes towards such tools change over time?* My evaluation of the SPECTRA prototype explored a single training cycle in users' homes, and the next step is a field deployment and longitudinal study to understand how users adapt to these tools in their everyday routines and whether their engagement with the model's training changes their feelings about these tools [60]. To match the format and feedback preferences of most DHH users [37, 46], this system should consist of a smartphone-based training workflow (following design guidelines from 5.5) in tandem with a smartwatch-based listener (*i.e.*, for sampling and model deployment).

This research should also explore how DHH users choose to refine their models over time and across many iterative training cycles. *How do DHH users approach training and performance optimization as an ongoing process?* Do they continually collect and adjust training data, or do they reach a point of satisfactory performance where further effort is deemed unnecessary? *Further, how do users decide when a model's performance is "good enough"?* Traditional evaluation measures (*e.g.*, F-scores) may not provide practical meaning about the system's uncertainty to end users. Instead, future work may look to Kay *et al.*'s framework for reaching acceptable accuracy [77] by capturing DHH users' ratings on hypothetical sound recognition scenarios to identify their preferred relative weights for precision versus recall, then adjusting the model accordingly. Investigating these and related questions would yield greater insight into the long-term sustainability of these tools.

Finally, *how does context influence the way DHH users train and refine their models?* For example, some users might train a new model for different sounds in each of their intended contexts of use (*e.g.*, at home, at work, while outdoors), while those with fewer sound interests might opt for a single, more general model. Jain *et al.*'s ProtoSound architecture [67] offers flexibility for varied contexts if desired, and understanding how contextual factors affect users training choices can further inform the design of adaptable tools.

Optimized Training Process. Allowing users to engage with the ML pipeline interactively can provide a sense of transparency and control [29], which can positively impact their trust, satisfaction, and long-term use of ML systems [3, 83]. However, my work shows that a lengthy training workflow can lead to fatigue and disengagement after just one training cycle. *How can we optimize the training process to reduce effort while still fostering thoughtful engagement from users?* SPECTRA's highly granular workflow may still hold value as an optional step when deeper troubleshooting is needed or as an educational tool for DHH people who are interested in learning more about ML within a highly practical application. However, future work should compare workflows of decreasing interaction complexity (with SPECTRA at the maximum) to search for the optimal level of time investment *vs.* active, thoughtful participation and learning among users.

My work found that clustering visualizations can be valuable tools for DHH users to understand audio datasets and refine training data. However, some users may not want to collect data themselves or train a fully custom model. *Can we incorporate clustering into existing sound recognition applications to support DHH users' understanding of these models' strengths and weaknesses?* For example, pre-trained models (*e.g.*, HomeSound [71], iPhone Sound Notifications) could passively save sound samples over several days (similar to Wu *et al.*'s ListenLearner [146]), then apply unsupervised clustering to this large dataset within an embedding generated from the underlying model. Users could explore the character of their soundscape by how their model tends to group or differentiate this dataset.

Clustering of passively sampled datasets is also a promising direction for personalization requiring minimal data collection effort from users. For models that are extensible to new sound classes (*e.g.*, Android [51]).

Users could assess the potential performance of new sound classes based on the proximity of that sound's embedded samples to the rest of the soundscape population. For example, a user may find the off-the-shelf model is inconvenient at mealtime due to classifying both their rice cooker and microwave as "Appliance beep", despite each cluster being well-separated from each other (and the rest of the dataset)—prompting them to create separate sound classes for both appliances.

Privacy Considerations. Training a personalized sound recognizer requires users to record samples of sounds in their environment, but during this process, DHH users may not realize when they have captured sensitive information. *How can we design tools that effectively protect the privacy of DHH users and those around them?* While my evaluation of wearable sound feedback found few privacy concerns among DHH users towards wearable devices with always-on audio sensing, further research is needed to understand their perspectives on capturing and storing personal audio datasets. This line of inquiry builds upon the work of Stangl *et al.* exploring the concerns of people who are blind to personal visual content [130, 131]. Personalizable tools should prioritize privacy-preserving techniques, such as storing only non-reconstructable audio features [72].

In addition to data privacy, *what are users' social and privacy considerations with recording audio in public spaces?* While smartwatch-based sound feedback was perceived to be discreet, recording audio examples with a phone or watch may include more overt signs of sensing and data capture, potentially raising concerns among bystanders. Communicating the assistive nature of the device can mitigate some negative attitudes towards sensing technology [115], but individuals with disabilities may also be sensitive to devices that draw unwanted attention to their disability [125]. Unfortunately, my field study of recording practices was conducted in 2020 at the height of the COVID-19 pandemic, when most participants could not record in public contexts; future longitudinal studies should investigate this topic further.

6.4 Limitations

I outline the core limitations of this dissertation related to the scope of my research methodology, possibilities for interactive ML, and focus on individual experiences. I also provide avenues for addressing them in further future work.

Lack of Performance Evaluations. My dissertation employed an empirical, qualitative research methodology centered on understanding user experiences, challenges, and design considerations for accessible sound recognition tools. This approach yielded crucial insights into accessibility needs, individual preferences, and technological possibilities to guide the design of future systems in this space. However, this methodological focus on understanding the subjective aspects of user experience did not provide more generalizable performance metrics or allow for direct comparison of design features based on objective measures. For instance, my investigation of sound feedback did not measure participants' reaction time to tactons or their accuracy in locating sound sources from directional feedback. Similarly, my evaluation of the SPECTRA prototype did not compare how models trained on datasets curated using SPECTRA's visualizations performed against those trained on unrefined sampling datasets. Future research incorporating quantitative performance evaluations could complement these qualitative insights to provide a more comprehensive understanding of the relationship between DHH user interaction, design features, and models' performance.

Limited Scope of Interactive ML Possibilities. The scope of my work was limited to supervised learning and batch training processes, which have become standard for the training of Convolutional Neural Network (CNN)-based sound models [55], but only capture only a subset of the broader interactive ML landscape [152]. While this scope allowed exploration of a personalization workflow that aligns with popular sound recognition tools, future research could explore interactive training incorporated into other ML methods, such as unsupervised [146] or reinforcement learning [28]. Future architectures may also support sequential training of sound models (*e.g., [34, 122]*, which may offer greater cause-and-effect insight through the ordering of training data. In addition, my work focused on users adjusting their models via dataset creation and curation, but future work could investigate more granular model adjustments (*e.g.*, adjusting weights, thresholds [30]), which may result in further performance improvements.

Individualized Scope. Finally, I want to acknowledge my research's focus on the individual use of sound recognition tools as a limitation in itself. I carried out my work with the belief that sound recognition tools should support *independent* personalization as a baseline for DHH users, but some may still feel unqualified for this task. Exploring collaborative personalization scenarios, where DHH users record, train, and deploy with assistance from friends or family members—suggested by a few participants in my work—could highlight potential strengths of *interdependent* usage of these tools. Future work should also consider the broader social context of personalizable sound recognition tools and explore their potential impact on interactions between and among DHH individuals, such as for facilitating communication (*e.g.*, name calls) or social inclusion, as well as potential issues related to privacy.

6.5 Closing Remark: Towards Transparency in Accessible AI

AI systems, particularly generative tools like ChatGPT and DALL-E, are steadily being incorporated into our lives, yet in most cases, they remain a "black box" to end-users. When users cannot understand how AI systems make decisions, they may be reluctant to place their trust in them [29]. This lack of transparency is especially problematic for sound recognition tools, where users hope the system will provide accurate and reliable information about their environment. When a black box system makes an error, it may be impossible to identify the cause of that error; in accessibility applications where the input itself is inaccessible, users might not be able to verify the system's output at all.

My work advocates for greater transparency and control in all AI systems, particularly those designed for accessibility. Sound recognition tools are widely available, but many DHH users remain dissatisfied with their performance [60, 67]. Building interactive AI systems that allow users to explore decision-making processes can reveal strengths and limitations, foster trust, and allow users to anticipate and safeguard against potential errors. Further, granular control over AI's learning enables personalization, allowing users to tailor the system to their specific needs and preferences. The risk of biased output [145] can be mitigated in systems trained on or optimized for end-user data, increasing equitable access for all.

However, we must recognize that even the most accurate, transparent, and controllable implementation of these tools cannot replace hearing or "solve" disability. A more accessible world requires greater accommodations and changes throughout society, such as hearing individuals taking the initiative in resolving access burdens (*e.g.*, [100]). As my work shows, DHH individuals have diverse preferences and ways of navigating the world, and some may prefer non-auditory modes of living. For those who desire greater access to sound information, however, sound recognition tools should be built to augment, not replace, their existing strategies and accommodations. My work offers an initial step toward building human-centered sound recognition tools. Continuing this work in partnership with the DHH community—with transparency and control as central tenets—can bring forth a future of AI for accessibility that is designed on the terms of those it intends to serve.

References

[1] Dustin Adams, Tory Gallagher, Alexander Ambard, and Sri Kurniawan. 2013. Interviewing blind photographers: design insights for a smartphone application. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 1–2. https://doi.org/10.1145/2513383.2513418

[2] A. Akbari and R. Jafari. 2020. Personalizing Activity Recognition Models Through Quantifying Different Types of Uncertainty Using Wearable Sensors. *IEEE Transactions on Biomedical Engineering* 67, 9 (2020), 2530–2541. https://doi.org/10.1109/TBME.2019.2963816

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105. https://doi.org/10.1609/aimag.v35i4.2513

[4] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 337–346. https://doi.org/10.1145/2702123.2702509

[5] Erin E. Andrews. 2017. Disability Models. In *Practical Psychology in Medical Rehabilitation*, Maggi A. Budd, Sigmund Hough, Stephen T. Wegener, and William Stiers (Eds.). Springer International Publishing, Cham, 77–83. https://doi.org/10.1007/978-3-319-34034-0_9

[6] Apple. [n. d.]. iOS 14 is available today. https://www.apple.com/newsroom/2020/09/ios-14-is-available-today/

[7] Apple. 2023. Recognize sounds using iPhone. https://support.apple.com/guide/iphone/use-sound-recognition-iphf2dc33312/17.0/ios/17.0

[8] Audacity Team. 2020. Audacity(R): Free Audio Editor and Recorder. https://audacityteam.org/

[9] Juhee Bae, Tove Helldin, Maria Riveiro, Sławomir Nowaczyk, Mohamed-Rafik Bouguelia, and Göran Falkman. 2020. Interactive Clustering: A Comprehensive Review. *Comput. Surveys* 53, 1 (Feb. 2020), 1:1–1:39. https://doi.org/10.1145/3340960

[10] Thomas Balkany, Annelle V Hodges, and Kenneth W Goodman. 1996. Ethics of cochlear implantation in young children. *Otolaryngology—Head and Neck Surgery* 114, 6 (1996), 748–755. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[11] H-Dirksen L. Bauman and Joseph J. Murray. 2010. *Deaf Studies in the 21st Century: "Deaf-gain" and the Future of Human Diversity*. Vol. 2. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195390032.013.0014 Publication Title: The Oxford Handbook of Deaf Studies, Language, and Education.

[12] Tanja Blascheck, Lonni Besancon, Anastasia Bezerianos, Bongshin Lee, and Petra Isenberg. 2019. Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 630–640. https://doi.org/10.1109/TVCG.2018.2865142

[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 (ASSETS '16)*. ACM Press, New York, New York, USA, 3–13. https://doi.org/10.1145/2982142.2982171

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

[15] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[16] Stephen Brewster and Lorna M Brown. 2004. Tactons: Structured Tactile Messages for Non-visual Information Display. In *Proceedings of the Fifth Conference on Australasian User Interface - Volume 28 (AUIC '04)*. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 15–23.

[17] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382839

[18] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. 2017. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–21. https://doi.org/10.1145/3134664

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

[20] Himanshu Chaurasiya. 2020. Time-Frequency Representations: Spectrogram, Cochleogram and Correlogram. *Procedia Computer Science* 167 (2020), 1901–1910. https://doi.org/10.1016/j.procs.2020.03.209

[21] Yang Chen. 2017. Visualizing Large Time-series Data on Very Small Screens. In *EuroVis 2017 - Short Papers*, Barbora Kozlikova, Tobias Schreck, and Thomas Wischgoll (Eds.). The Eurographics Association. https://doi.org/10.2312/eurovisshort.20171130

[22] Naomi B. H. Croghan, Kathryn H. Arehart, and James M. Kates. 2014. Music Preferences With Hearing Aids. *Ear and Hearing* 35, 5 (2014), e170–e184. https://doi.org/10.1097/AUD.0000000000000056

[23] 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 *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*. IEEE, Las Vegas, NV, USA, 297–302. https://doi.org/10.1109/ICIS.2015.7166609

[24] Debanjan Datta and Gerald Friedland. 2023. Efficient Multimedia Computing: Unleashing the Power of AutoML. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 9700–9701. https://doi.org/10.1145/3581783.3613858

[25] Allan G. de Oliveira, Thiago M. Ventura, Todor D. Ganchev, Josiel M. de Figueiredo, Olaf Jahn, Marinez I. Marques, and Karl-L. Schuchmann. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics* 98 (Nov. 2015), 34–42. https://doi.org/10.1016/j.apacoust.2015.04.014

[26] Alex De Robertis and Ian Higginbottom. 2007. A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. *ICES Journal of Marine Science* 64, 6 (2007), 1282–1291. Publisher: Oxford University Press.

[27] Harvey Dillon. 2008. *Hearing aids.* Hodder Arnold.

[28] Hang Do, Quan Dang, Jeremy Zhengqi Huang, and Dhruv Jain. 2023. AdaptiveSound: An Interactive Feedback-Loop System to Improve Sound Recognition for Deaf and Hard of Hearing Users. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3597638.3608390

[29] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 297–307. https://doi.org/10.1145/3377325.3377501

[30] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (July 2018), 1–37. https://doi.org/10.1145/3185517

[31] Claire Edwards and Gill Harold. 2014. DeafSpace and the principles of universal design. *Disability and Rehabilitation* 36, 16 (Aug. 2014), 1350–1359. https://doi.org/10.3109/09638288.2014.913710

[32] 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 assessment 2005 : proceedings of the 3rd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.* University of Iowa, Iowa City, Iowa, 232–239. https://doi.org/10.17077/drivingassessment.1166

[33] Nirmala Erevelles and Andrea Minear. 2010. Unspeakable Offenses: Untangling Race and Disability in Discourses of Intersectionality. *Journal of Literary & Cultural Disability Studies* 4, 2 (Jan. 2010), 127–145. https://doi.org/10.3828/jlcds.2010.11

[34] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces - IUI '03.* ACM Press, New York, New York, USA, 39. https://doi.org/10.1145/604045.604056

[35] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11.* ACM Press, New York, New York, USA, 147. https://doi.org/10.1145/1978942.1978965

[36] Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the Special Issue on Human-Centered Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (July 2018), 1–7. https://doi.org/10.1145/3205942

[37] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300276

[38] Leah Findlater, Steven Goodman, Yuhang Zhao, Shiri Azenkot, and Margot Hanley. 2020. Fairness issues in AI systems that augment sensory abilities. *ACM SIGACCESS Accessibility and Computing* 125 (March 2020), 1–1. https://doi.org/10.1145/3386296.3386304 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[39] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 1126–1135.

[40] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound Datasets: a platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. Suzhou, China, 486–493.

[41] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: Composing Interactive Machine Learning Workflows and Interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 39–53. https://doi.org/10.1145/3472749.3474734

[42] Karyn L. Galvin, Jan Ginis, Robert S. C. 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 (May 2001), 18–36. https://doi.org/10.1375/audi.23.1.18.31095

[43] Rosemarie Garland-Thomson and Paul K. Longmore. 2003. Statement of Principles. https://dsq-sds.org/ Publication Title: Disability Studies Quarterly.

[44] 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)*. IEEE, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[45] L. Giuliani, L. Brayda, S. Sansalone, S. Repetto, and M. Ricchetti. 2017. Evaluation of a complementary hearing aid for spatial sound segregation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 221–225. https://doi.org/10.1109/ICASSP.2017.7952150

[46] 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 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376406

[47] 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 (June 2021), 1–23. https://doi.org/10.1145/3463501

[48] Steven M. Goodman, Emma J. McDonnell, Jon E. Froehlich, and Leah Findlater. 2024. SPECTRA: Personalizable Sound Recognition for Deaf and Hard of Hearing Users through Interactive Machine Learning. (2024).

[49] Google. 2020. Audio Model - Teachable Machines. https://teachablemachine.withgoogle.com/train/audio

[50] Google. 2020. Important household sounds become more accessible. https://blog.google/products/android/new-sound-notifications-on-android/

[51] Google. 2022. Get more done and have fun with new Android features. https://blog.google/products/android/new-android-features-september-2022/

[52] 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 (ASSETS '14).* ACM Press, New York, New York, USA, 337–338. https://doi.org/10.1145/2661334.2661410

[53] J. Harkins, P. E. Tucker, N. Williams, and J. Sauro. 2010. Vibration Signaling in Mobile Devices for Emergency Alerting: A Study With Deaf Evaluators. *Journal of Deaf Studies and Deaf Education* 15, 4 (Oct. 2010), 438–445. https://doi.org/10.1093/deafed/enq018

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. _eprint: 1512.03385.

[55] 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).* IEEE, 131–135.

[56] F. Wai-ling Ho-Ching, Jennifer Mankoff, and James A. Landay. 2003. Can You See What I Hear?: The Design and Evaluation of a Peripheral Sound Display for the Deaf. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03).* ACM Press, New York, New York, USA, 161–168. https://doi.org/10.1145/642611.642641

[57] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376177

[58] Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zamiri Zeraati, Ebrima Jarjue, Kyungjun Lee, and Hernisa Kacorri. 2022. Blind Users Accessing Their Training Images in Teachable Object Recognizers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility.* ACM, Athens Greece, 1–18. https://doi.org/10.1145/3517428.3544824

[59] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the Perception of Machine Teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376428

[60] Jeremy Zhengqi Huang, Hriday Chhabria, and Dhruv Jain. 2023. "Not There Yet": Feasibility and Challenges of Mobile Sound Recognition to Support Deaf and Hard-of-Hearing People. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York NY USA, 1–14. https://doi.org/10.1145/3597638.3608431

[61] Tom Humphries. 1977. Communicating Across Cultures (Deaf/Hearing) and Language Learning. (1977).

[62] Hilary Hutchinson, Heiko Hansen, Nicolas Roussel, Björn Eiderbäck, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, and Helen Evans. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the conference on Human factors in computing systems - CHI '03*. ACM Press, New York, New York, USA, 17. https://doi.org/10.1145/642611.642616

[63] Tatsuya Ishibashi, Yuri Nakao, and Yusuke Sugano. 2020. Investigating audio data visualization for interactive sound recognition. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 67–77. https://doi.org/10.1145/3377325.3377483

[64] 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 (ISWC '20)*. Association for Computing Machinery, New York, NY, USA, 55–57. https://doi.org/10.1145/3410531.3414291 event-place: Virtual Event, Mexico.

[65] Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E Froehlich. 2015. Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing. In *Proceedings of the 33rd Annual ACM Conference*

on Human Factors in Computing Systems (CHI '15). ACM, New York, NY, USA, 241–250. https://doi.org/10.1145/2702123.2702393

[66] 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 *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '18)*. ACM, New York, NY, USA, 81–92. https://doi.org/10.1145/3234695.3236362

[67] Dhruv Jain, Khoa Huynh Anh Nguyen, Steven M. Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon E. Froehlich. 2022. ProtoSound: A Personalized and Scalable Sound Recognition System for Deaf and Hard-of-Hearing Users. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3502020

[68] 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*. ACM, Virtual Event USA, 160–170. https://doi.org/10.1145/3461778.3462106

[69] 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 *Proceedings of the 2021 International Conference on Multimodal Interaction*. ACM, Montréal QC Canada, 80–91. https://doi.org/10.1145/3462244.3479946

[70] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, 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 (CHI '19)*. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300324

[71] 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*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376758

[72] 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 *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2020)*. ACM.

[73] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '11*. ACM Press, New York, New York, USA, 203. https://doi.org/10.1145/2049536.2049573

[74] Hernisa Kacorri. 2017. Teachable Machines for Accessibility. *SIGACCESS Access. Comput.* 119 (Nov. 2017), 10–18. https://doi.org/10.1145/3167902.3167904 Place: New York, NY, USA Publisher: ACM.

[75] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5839–5849. https://doi.org/10.1145/3025453.3025899

[76] Yoshihiro Kaneko, Inho Chung, and Kenji Suzuki. 2013. Light-Emitting Device for Supporting Auditory Awareness of Hearing-Impaired People during Group Conversations. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 3567–3572. https://doi.org/10.1109/SMC.2013.608 Backup Publisher: IEEE.

[77] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 347–356. https://doi.org/10.1145/2702123.2702603

[78] Bongjun Kim and Bryan Pardo. 2017. I-SED: an Interactive Sound Event Detector. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 553–557. https://doi.org/10.1145/3025171.3025231

[79] Bongjun Kim and Bryan Pardo. 2018. A Human-in-the-Loop System for Sound Event Detection and Annotation. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (July 2018), 1–23. https://doi.org/10.1145/3214366

[80] Ki-Won Kim, Jung-Woo Choi, and Yang-Hann Kim. 2013. An Assistive Device for Direction Estimation of a Sound Source. *Assistive Technology* 25, 4 (Oct. 2013), 216–221. https://doi.org/10.1080/10400435.2013.768718

[81] W. Bradley Knox and Peter Stone. 2015. Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. *Artificial Intelligence* 225 (Aug. 2015). http://www.cs.utexas.edu/users/ai-lab?knox:aij15

[82] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[83] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. ACM Press, New York, New York, USA, 1. https://doi.org/10.1145/2207676.2207678

[84] Raja Kushalnagar. 2019. Deafness and Hearing Loss. In *Web Accessibility*. Springer, London, UK, 35–47. https://doi.org/10.1007/978-1-4471-7440-0_3

[85] Paddy Ladd. 2003. *Understanding deaf culture: In search of deafhood*. Multilingual Matters.

[86] Paddy Ladd and Harlan Lane. 2013. Deaf Ethnicity, Deafhood, and Their Relationship. *Sign Language Studies* 13, 4 (2013), 565–579. https://doi.org/10.1353/sls.2013.0012

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

[88] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting Blind Photography in the Context of Teachable Object Recognizers. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. ACM, New York, NY, USA, 83–95. https://doi.org/10.1145/3308561.3353799

[89] Seungyon "Claire" Lee and Thad Starner. 2010. BuzzWear: Alert Perception in Wearable Tactile Displays on the Wrist. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 433–442. https://doi.org/10.1145/1753326.1753392

[90] Ziming Li, Shannon Connell, Wendy Dannels, and Roshan Peiris. 2022. SoundVizVR: Sound Indicators for Accessible Sounds in Virtual Reality for Deaf or Hard-of-Hearing Users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Athens Greece, 1–13. https://doi.org/10.1145/3517428.3544817

[91] 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 (Oct. 2002), 504–516. https://doi.org/10.1109/TSA.2002.804546

[92] Deborah Lupton and Wendy Seymour. 2000. Technology, selfhood and physical disability. *Social Science & Medicine* 50, 12 (2000), 1851–1862. https://doi.org/10.1016/S0277-9536(99)00422-0

[93] Makeability Lab. 2020. SoundWatch. https://github.com/makeabilitylab/SoundWatch

[94] Shoji Makino, Shoko Araki, Ryo Mukai, and Hiroshi Sawada. 2004. Audio source separation based on independent component analysis. In *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No. 04CH37512)*, Vol. 5. IEEE, V–V.

[95] Jennifer Mankoff, Gillian R. Hayes, and Devva Kasnitz. 2010. Disability studies as a source of critical inquiry for the field of assistive technology. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '10*. ACM Press, New York, New York, USA, 3. https://doi.org/10.1145/1878803.1878807

[96] Patrizia Marti, Michele Tittarelli, Matteo Sirizzotti, Iolanda Iacono, and Riccardo Zambon. 2019. From Stigma to Objects of Desire: Participatory Design of Interactive Jewellery for Deaf Women. In *Interactivity, Game Creation, Design, Learning, and Innovation*. Springer, 429–438. https://doi.org/10.1007/978-3-030-06134-0_46

[97] 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 the 8th International Conference on Ubiquitous Computing (UbiComp '06)*. Springer-Verlag, 159–176. https://doi.org/10.1007/11853565_10

[98] 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 (July 2006), 333–351. https://doi.org/10.1080/01449290600636488

[99] Tara Matthews, Janette Fong, and Jennifer Mankoff. 2005. Visualizing non-speech sounds for the deaf. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '05)*. ACM Press, New York, New York, USA, 52. https://doi.org/10.1145/1090785.1090797

[100] Emma McDonnell. 2022. Understanding Social and Environmental Factors to Enable Collective Access Approaches to the Design of Captioning Technology. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3517428.3550417

[101] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426 arXiv:1802.03426 [cs, stat].

[102] Matthias Mielke and Rainer Bruck. 2016. AUDIS wear: A smartwatch based assistive device for ubiquitous awareness of environmental sounds. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5343–5347. https://doi.org/10.1109/EMBC.2016.7591934

[103] Matthias Mielke and Rainer Brück. 2015. Design and evaluation of a smartphone application for non-speech sound awareness for people with hearing loss. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5008–5011. https://doi.org/10.1109/EMBC.2015.7319516

[104] 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 (ASSETS '15)*. ACM Press, New York, New York, USA, 301–302. https://doi.org/10.1145/2700648.2811347

[105] Matthew S. Moore and Linda Levitan. 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, Rochester, NY, USA.

[106] Meredith Ringel Morris. 2020. AI and Accessibility: A Discussion of Ethical Considerations. *Commun. ACM* (June 2020). https://www.microsoft.com/en-us/research/publication/ai-and-accessibility-a-discussion-of-ethical-considerations/

[107] Yuri Nakao and Yusuke Sugano. 2020. Use of Machine Learning by Non-Expert DHH People: Technological Understanding and Sound Perception. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3419249.3420157

[108] Michael Oliver. 1990. New Social Movements. In *The Politics of Disablement*. Macmillan Education UK, London, 112–131. https://doi.org/10.1007/978-1-349-20895-1_8

[109] Yi-Hao Peng, Ming-Wei Hsi, 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 *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 293:1–293:10. https://doi.org/10.1145/3173574.3173867

[110] Erik Pescara, Alexander Wolpert, Matthias Budde, Andrea Schankin, and Michael Beigl. 2017. Lifetact: Utilizing Smartwatches As Tactile Heartbeat Displays in Video Games. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia (MUM '17)*. ACM, New York, NY, USA, 97–101. https://doi.org/10.1145/3152832.3152863

[111] A. J. Phillips, A. R. D. Thornton, S. Worsfold, A. Downie, and J. Milligan. 1994. Experience of using vibrotactile aids with the profoundly deafened. *International Journal of Language & Communication Disorders* 29, 1 (Jan. 1994), 17–26. https://doi.org/10.3109/13682829409041478

[112] Martin Pielot, Benjamin Poppinga, and Susanne Boll. 2010. PocketNavigator. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services - MobileHCI '10 (MobileHCI '10)*. ACM Press, New York, New York, USA, 423. https://doi.org/10.1145/1851600.1851696

[113] Meg Pirrung, Nathan Hilliard, Artëm Yankov, Nancy O'Brien, Paul Weidert, Courtney D Corley, and Nathan O Hodas. 2018. Sharkzor: Interactive Deep Learning for Image Triage, Sort and Summary. _eprint: 1802.05316.

[114] Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. 2016. Smartwatch in vivo. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 5456–5469. https://doi.org/10.1145/2858036.2858522

[115] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings*

of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). ACM, New York, NY, USA, 4884–4895. https://doi.org/10.1145/2858036.2858130

[116] Thejan Rajapakshe, Rajib Rana, Siddique Latif, Sara Khalifa, and Björn W. Schuller. 2019. Pre-training in Deep Reinforcement Learning for Automatic Speech Recognition. _eprint: 1910.11256.

[117] 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 (Nov. 2020), 413–451. https://doi.org/10.1080/07370024.2020.1734931

[118] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging Perspectives in Human-Centered Machine Learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299014

[119] Rev.com. 2020. Voice Recorder App | Audio Recording App. https://www.rev.com/voicerecorder

[120] James Robert, Marc Webbie, and others. 2018. Pydub. http://pydub.com/

[121] Stefano Rovetta, Zied Mnasri, and Francesco Masulli. 2020. *Detection of Hazardous Road Events From Audio Streams: An Ensemble Outlier Detection Approach*. https://doi.org/10.1109/EAIS48028.2020.9122704 Pages: 6.

[122] Téo Sanchez, Baptiste Caramiaux, Jules Françoise, Frédéric Bevilacqua, and Wendy E. Mackay. 2021. How do People Train a Machine? Strategies and (Mis)Understandings. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 162:1–162:26. https://doi.org/10.1145/3449236

[123] Caitlyn Seim, Rodrigo Pontes, Sanjana Kadiveti, Zaeem Adamjee, Annette Cochran, Timothy Aveni, Peter Presti, and Thad Starner. 2018. Towards Haptic Learning on a Smartwatch. In *Proceedings of the*

2018 ACM International Symposium on Wearable Computers (ISWC '18). ACM, New York, NY, USA, 228–229. https://doi.org/10.1145/3267242.3267269

[124] Tom Shakespeare. 2010. *The Social Model of Disability* (3 ed.). Routledge. Publication Title: The Disability Studies Reader Section: 16.

[125] Kristen Shinohara and Jacob O Wobbrock. 2011. In the Shadow of Misperception: Assistive Technology Use and Social Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 705–714. https://doi.org/10.1145/1978942.1979044

[126] Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. 2017. Active learning for sound event classification by clustering unlabeled data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 751–755. https://doi.org/10.1109/ICASSP.2017.7952256

[127] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (June 2017), 17:1–17:21. https://doi.org/10.1145/3090082 Place: New York, NY, USA Publisher: ACM.

[128] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[129] Joan Sosa-García and Francesca Odone. 2017. "Hands On" Visual Recognition for Visually Impaired Users. *ACM Transactions on Accessible Computing* 10, 3 (Aug. 2017), 1–30. https://doi.org/10.1145/3060056

[130] Abigale Stangl, Kristina Shiroma, Nathan Davis, Bo Xie, Kenneth R. Fleischmann, Leah Findlater, and Danna Gurari. 2022. Privacy Concerns for Visual Assistance Technologies. *ACM Transactions on Accessible Computing* 15, 2 (May 2022), 15:1–15:43. https://doi.org/10.1145/3517384

[131] Abigale Stangl, Kristina Shiroma, Bo Xie, Kenneth R. Fleischmann, and Danna Gurari. 2020. Visual Content Considered Private by People Who are Blind. In *Proceedings of the 22nd International ACM*

SIGACCESS Conference on Computers and Accessibility (ASSETS '20). Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3373625.3417014

[132] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. 2020. AnchorViz: Facilitating Semantic Data Exploration for IML. *ACM Transactions on Interactive Intelligent Systems* 10, 1 (Jan. 2020), 1–38. https://doi.org/10.1145/3241379

[133] 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 (Jan. 1981), 195–199. https://doi.org/10.3109/03005368109081437

[134] Kyle A. Swiston and Daniel J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* 80, 1 (March 2009), 42–50. https://doi.org/10.1111/j.1557-9263.2009.00204.x

[135] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A Survey on Deep Transfer Learning. _eprint: 1808.01974.

[136] TensorFlow. [n. d.]. Pre-trained TensorFlow.js models. https://github.com/tensorflow/tfjs-models/

[137] Martin Tomitsch and Thomas Grechenig. 2007. Design Implications for a Ubiquitous Ambient Sound Display for the Deaf.. In *Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Assistive Technology for All Ages (CVHI 2007)*. M.A. Hersh (Ed.).

[138] Bonnie Poitras Tucker. 1998. Deaf Culture, Cochlear Implants, and Elective Disability. *The Hastings Center Report* 28, 4 (July 1998), 6. https://doi.org/10.2307/3528607

[139] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-tech. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. ACM Press, New York, New York, USA, 31–40. https://doi.org/10.1145/1240624.1240630

[140] Donald A Vogel, Patricia A McCARTHY, Gene W Bratt, and Carmen Brewer. 2007. The clinical audiogram: its history and current use. *Commun Disord Rev* 1, 2 (2007), 81–94.

[141] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '12*. ACM Press, New York, New York, USA, 95. https://doi.org/10.1145/2384916.2384934

[142] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. 2019. Using Expert Patterns in Assisted Interactive Machine Learning: A Study in Machine Teaching. 578–599. https://doi.org/10.1007/978-3-030-29387-1_34

[143] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J. Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. ATMSeer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300911

[144] Pete Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. https://doi.org/10.48550/arXiv.1804.03209 arXiv:1804.03209 [cs].

[145] Jason J. G. White. 2022. Artificial Intelligence and People with Disabilities: a Reflection on Human–AI Partnerships. In *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*, Fang Chen and Jianlong Zhou (Eds.). Springer International Publishing, Cham, 279–310. https://doi.org/10.1007/978-3-030-72188-6_14

[146] 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376875

[147] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. 2021. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association

for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445306

[148] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, New York, NY, USA, 573–584. https://doi.org/10.1145/3196709.3196729

[149] Eddy Yeung, Arthur Boothroyd, and Cecil Redmond. 1988. A wearable multichannel tactile display of voice fundamental frequency. *Ear and hearing* 9, 6 (1988), 342–350.

[150] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? http://arxiv.org/abs/1411.1792 arXiv:1411.1792 [cs].

[151] 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 (Aug. 2005), 1003–1015. https://doi.org/10.1121/1.1945787

[152] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. 2018. An Overview of Machine Teaching. http://arxiv.org/abs/1801.05927 arXiv:1801.05927 [cs].