# Toward User-Driven Sound Recognizer Personalization With People Who Are d/Deaf or Hard of Hearing

STEVEN M. GOODMAN, University of Washington, USA
PING LIU, University of Washington, USA
DHRUV JAIN, University of Washington, USA
EMMA J. MCDONNELL, University of Washington, USA
JON E. FROEHLICH, University of Washington, USA
LEAH FINDLATER, University of Washington, USA



Automated sound recognition tools can be a useful complement to d/Deaf and hard of hearing (DHH) people's typical communication and environmental awareness strategies. Pre-trained sound recognition models, however, may not meet the diverse needs of individual DHH users. While approaches from human-centered machine learning can enable non-expert users to build their own automated systems, end-user ML solutions that augment human sensory abilities present a unique challenge for users who have sensory disabilities: how can a DHH user, who has difficulty hearing a sound themselves, effectively record samples to train an ML system to recognize that sound? To better understand how DHH users can drive personalization of their own assistive sound recognition tools, we conducted a three-part study with 14 DHH participants: (1) an initial interview and demo of a personalizable sound recognizer, (2) a week-long field study of in situ recording, and (3) a follow-up interview and ideation session. Our results highlight a positive subjective experience when recording and interpreting training data in situ, but we uncover several key pitfalls unique to DHH users—such as inhibited judgement of representative samples due to limited audiological experience. We share implications of these results for the design of recording interfaces and human-the-the-loop systems that can support DHH users to build sound recognizers for their personal needs.

CCS Concepts: • **Human-centered computing** → *Empirical studies in accessibility*; *Accessibility technologies.*

Additional Key Words and Phrases: Deaf and hard of hearing, accessibility, sound recognition, field study

**ACM Reference Format:**
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 63 (June 2021), 23 pages. https://doi.org/10.1145/3463501

## 1 INTRODUCTION

Recent advances in machine learning (ML) and signal processing have enabled new automatic sound recognition tools for d/Deaf and hard of hearing (DHH) users. Work in sound awareness [8, 27, 37, 53, 54] shows that DHH

Authors' addresses: Steven M. Goodman, smgoodmn@uw.edu, University of Washington, Seattle, Washington, USA; Ping Liu, pl92@uw.edu, University of Washington, Seattle, Washington, USA; Dhruv Jain, djain@cs.washington.edu, University of Washington, Seattle, Washington, USA; Emma J. McDonnell, ejm249@uw.edu, University of Washington, Seattle, Washington, USA; Jon E. Froehlich, jonf@cs.washington.edu, University of Washington, Seattle, Washington, USA; Leah Findlater, leahkf@uw.edu, University of Washington, Seattle, Washington, USA.


2474-9567/2021/6-ART63 $15.00
https://doi.org/10.1145/3463501

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 [5] can notify users when it recognizes eleven sound categories (*e.g.,* baby crying, car horn), while Android's Sound Notifications feature [29] supports ten sounds plus a timeline of all recently detected sounds. However, these features—and prior work implementing sound classification for DHH users [37, 38, 55, 67]—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.

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 [8, 37]) 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 [21, 62] 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 [23, 40, 58]: 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.,* [41, 49, 69]) 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, 39, 74]) 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.* [8], 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.* [58] 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 [50]—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 [28];
(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 on 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 [58, 78]—but limited audiological 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 paper 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 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.

## 2 RELATED WORK

We review past research about sound awareness preferences of DHH people, tools that can support sound awareness, and human-centered machine learning.

### 2.1 Diversity of Sound Awareness Needs

While prior work has consistently found DHH participants to be interested in sound awareness tools, this interest is not uniform and is often conditioned on cultural identity and context. A DHH person may belong to Deaf (capital 'D'), deaf, or hard of hearing communities [13, 57]. Individuals who identify as Deaf follow an established set of norms, behaviors and language (called 'Deaf culture' [47, 57]). In contrast, for hard of hearing or deaf individuals, deafness is primarily an audiological experience [57]. The cultural difference may influence sound awareness needs. Bragg *et al.* [8] and Findlater *et al.* [22] conducted online surveys with 87 and 201 DHH participants, respectively, finding that hard of hearing users may be more interested in certain sounds (*e.g.*, phone ringing, spoken conversations) than users identifying as d/Deaf.

While accounting for the diverse perspectives of DHH people, prior work also highlights several general preferences among DHH users. Overall, the most desired sound characteristic is identity, which users prioritize when compared to other characteristics like volume or duration [8, 22, 27, 54]. When discussing sounds of interest, DHH users generally rank awareness of urgent sounds (*e.g.*, safety-related alarms, sirens) as most important, followed by sounds that indicate others' presence (*e.g.*, door knocks, footsteps) and appliance alerts (*e.g.*, oven timers, pop-up toasters) [8, 22, 54, 67]. Additionally, the relevance of sound information may change as the user moves between social contexts (*e.g.*, family *vs.* strangers) [8, 22, 36] and physical locations (*e.g.*, at home *vs.* while mobile) [27, 54]. For example, in the home, sound identity may be adequate [37], while directional indicators are important when mobile [27]. DHH users have frequently expressed interest in sounds that are specific to their life (*e.g.*, babies and children [54], name calls [8]) and fine-tuned sounds in their home [37], which points to the need for end users to be able to personalize their sound awareness tools—the focus of our study.

### 2.2 Sound Awareness Tools

Though our work does not contribute a new sound awareness tool, we turn to prior work in this domain to inform our study. An early project by Matthews *et al.* [53] examined a PDA-based prototype for DHH users to request human transcription of both speech and non-speech sounds in the most recent 30 seconds of audio, and was well-received despite misjudging relevant sounds. More recent work has aimed to provide broad sound recognition support by employing pre-trained classification models [37, 38, 56, 67]. For example, Sicong *et al.* [67] deployed a smartphone tool trained for nine sound classes—including police sirens and door knocks—to 86 participants for two days in a school setting. Users were satisfied with the tool, although some were concerned with the accuracy of short sound events (*e.g.*, coughing, crying). Jain *et al.* [37] installed a tablet-based sound classification system for 19 sounds in four homes, observing concerns over inconsistent classification and a desire to personalize the system for sounds specific to each home.

Personalization is essential to provide context-specific support and meet the wide-ranging needs of the DHH community [22, 27, 53, 55, 58]. Some projects have explored options for DHH users to filter notifications for certain sounds [27, 37, 38], but they stopped short of adding or modifying sound classes through user-provided recordings. Bragg *et al.* [8] designed a personalizable sound recognition smartphone app and recruited 12 DHH participants for a Wizard-of-Oz usability study. While participants found the recording and training workflow easy, they recorded samples of only two sounds (alarm clock, door knock) in an office setting—an experience that

is unlikely to represent the varied use-cases, sounds, and environmental noise in the daily life of DHH users. In contrast, we asked participants to select their own sound classes and record samples each day during a week-long field study to learn about practical aspects of the task.

## 2.3 Human-Centered Machine Learning

Human-centered machine learning research aims to design and build automated systems that can fulfill user goals, fit user-specific contexts, and accommodate people without programming experience [21, 62]. From this space, several approaches have emerged for people without ML expertise to build models of their own. With Automated Machine Learning (AutoML) approaches (*e.g.,* [77]), novice end-users provide a large batch of labeled data while traditional ML tasks such as feature engineering, model selection, and hyperparameter optimization are completed automatically [18]. In contrast to AutoML's black box approach to model creation, interactive machine learning (IML) leverages end-users as "humans-in-the-loop" to iteratively engage in building and refining ML models [3, 19, 61, 70]. An IML workflow involves a quick loop between use of the system and training of the target model [70], during which the user may provide indicative samples, describe salient features, or select high-level model parameters [19]. Interactive machine *teaching* [62] is a specific IML approach that takes the engagement further and positions the human-in-the-loop in the role of a teacher with rich knowledge of the task instead of a source for data labels [76]. However, work within the approaches above typically assumes that the end-user has domain expertise and can readily interact with the data underlying their intended model—an assumption that may not hold for DHH users and audio data.

In the field of accessibility, human-centered ML applications can allow disabled users to personalize data-driven assistive technology to meet their individual needs [40]. 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 [23, 40]. 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.*, [41, 69]) first examined via studies of users' needs in this context, and more recently began addressing through active feedback techniques to assist in image capture [49]. Indeed, in Nakao *et al.*'s [58] study of DHH users' technical understanding of ML, workshop participants struggled to choose acceptable sound samples for training due to a lack of non-auditory feedback. For work on sampling feedback, VoiceAssist [65] provides real-time visual feedback to help inexperienced users reduce reverberation and background noise in voice recordings, and a user study showed third party listeners preferred recordings made using VoiceAssist compared to those without. To our knowledge, however, prior work has yet to explore sound sampling feedback for sound recognition with any population—a gap which we begin to address with our work.

IML research for audio has primarily focused on sample annotation and labelling (*e.g.*, [30, 35, 42, 43, 66]). For interactive sound recognition, Ishibashi *et al.* [35] explored visualization options (*e.g.,* spectrograms, thumbnails) for browsing large sets of unlabelled audio samples via a clustering interface. Google's Teachable Machine experiment [11, 28] allows non-expert users to quickly train a personal sound recognition model with their own audio samples, but ML expertise is required to have agency over the produced model (*e.g.,* redefining features). Nakao *et al.* [58] studied non-expert DHH users' understanding of ML with a similar interactive workflow that supported training sets built from users' recordings or selected from a large sound library [26]. DHH participants overcame gaps in their technical understanding and identified additional use cases after hands-on experience, but some struggled to review samples and define ML tasks for sounds they knew but could not hear. This work, in combination with the work of Bragg *et al.* [8], begins to outline a design space for personalizable sound recognizers trained by non-expert DHH users. We continue this thread by focusing on how DHH users record and interpret audio samples for this purpose.

Table 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 |

## 3 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.

### 3.1 Participants

We recruited 14 DHH participants via email lists at two U.S. universities as well as via social media and snowball sampling (Table 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 and two used cochlear implants; one used both devices. 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.* [33], 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: use of a laptop or desktop computer at least once a week and use of 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.


**(a) Sound recognizer demo during the initial interview**

**(b) Sample recording during the field study**

**(c) Design brainstorming during the follow-up interview**

Fig. 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 [11, 28], 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.

## 3.2 Procedure

The study had three parts: an initial interview session to introduce audio recording for sound classification, one-week use of an audio recording application, and a final interview and design probe session (Figure 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 Zoom [80]. 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. We shared the interview materials in an online slide deck before the study (see Supplementary Materials) and employed Zoom'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 Zoom's chat feature.

Participants received non-auditory feedback via waveform and spectrogram sound visualizations during the initial session (Figure 2), and the waveform alone during the field study (Figure 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 [8]; 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 [16]), and can be difficult to interpret for novice hearing users [12, 35]. Early work showed frequency information was inadequate for DHH users in a sound identification task [54]; we briefly explore DHH participants' opinions of spectrograms for displaying sound activity. Below, we detail the three sessions of the study.

*3.2.1 Initial Session (75 min).* The initial session began with 15 minutes for Zoom 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 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 [11, 28] and its spectrogram visualization (Figure 1a). We led the same discussion with all participants during this activity but only ten of the 14 were able to record samples and train the model themselves The other four experienced technical difficulties and 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.,* "the 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 Zoom's annotation feature to explain Teachable Machine's interface but allowed participants to record samples on their own, as well as 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.[1] 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 transitioned to discussing possible characteristics of high quality sound samples for training a sound recognizer. Informed by training datasets used in prior work [37, 48], 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 [52] or remove background noise [17], we included both to our guide to prompt consideration of audiological 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 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 first, then the hearing first author connected salient areas of each visualizations 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.

*3.2.2 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 life for a week. At the end of the initial session, we helped participants download and configure the *Rev*

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

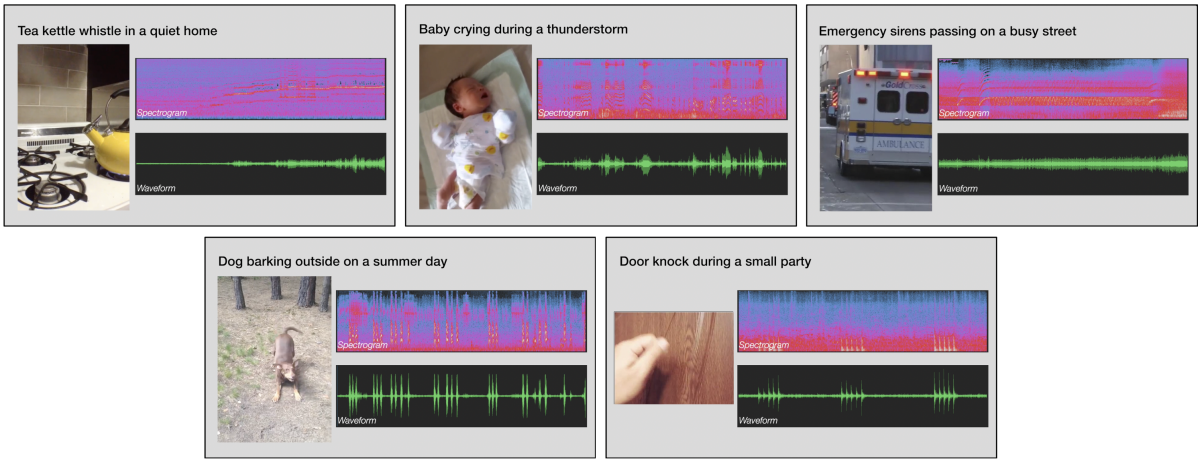


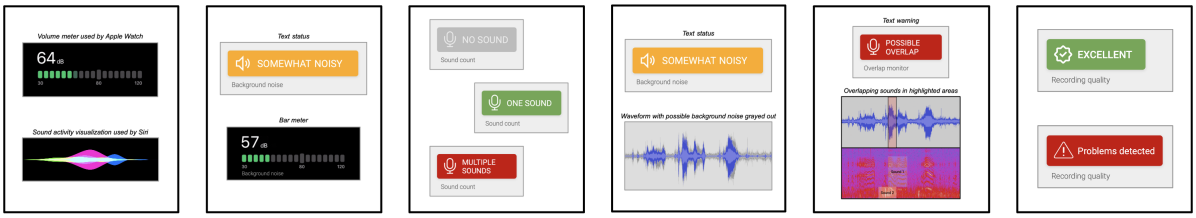

Fig. 2. Videos shown to participants to introduce real-life recording challenges and visualizations. To simulate a recording timeline, spectrogram and waveform visualizations were generated using Audacity [6] and set to advance in sync with the video clip. To match the described context of the *baby 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.

*Recorder* app [63] on their smartphone. In preparation for our study, we reviewed a variety of 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 disable this "auto-upload" function temporarily if needed—for instance, situations where 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 for flexibility did not provide a strict time limit. 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 those sounds, successes and challenges of recording, any sounds they had attempted but were unable to record, and any other information that might have helped with recording that day (see Supplementary Materials).

*3.2.3 Follow-up Interview and Design Probe Activity (60 min).* We scheduled a final Zoom video call during which the interviewer screen-shared a new slide deck (see Supplementary Materials). 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




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

Fig. 3. Six of ten sampling features taken from 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. For full context, see Supplementary Materials.

that and their initial expectations, and if they had changed their recording practices over the course of the week. Next, we provided a complete list of the sounds they had recorded and asked them to identify which had been the easiest and hardest to record as well as if they were satisfied with their recordings. Finally, we reviewed the list of high quality recording characteristics and discussed how each of the factors surfaced (if at all) during the week.

The second half of the session consisted of a design probe activity inspired by Hutchinson *et al.* [34] to discuss new ideas for supporting DHH people in independently sampling sounds. We first asked the participant to describe their ideal features, then presented ten possible feature ideas (Figure 3). For each idea, we showed a brief description and two mockups. We asked whether each feature would be useful and if the participant had any related design ideas of their own. 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.

### 3.3 Analysis and Positionality

Using reflexive thematic analysis [9, 10], 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, then applied these codes to data from two randomly selected participants. Another researcher reviewed the code applications, then met with the first author to discuss and refine the codes further. The first author coded the remaining transcripts, then 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. The first author—who ran all interviews and led analysis—is hearing. Some authors—who were involved in study design, analysis, and writing—are Deaf or hard of hearing. All members of the research team have backgrounds in human-computer interaction and many are computer scientists by training.

## 4 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 [78]): (1) data and label collection,

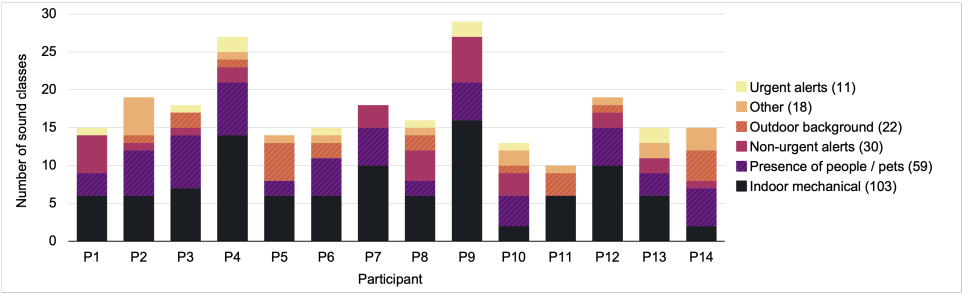



Fig. 4. Breakdown of sound categories recorded by participants. The total classes in each category are shown on the right.

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 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.

## 4.1 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) at an average 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 [64] to analyze each samples' 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 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"*.

## 4.2 ML Expertise and Prior Recording Experience

Participants demonstrated a range of ML knowledge in the first session despite all of them 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 [8, 22, 54]. 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 Planning and Recording Samples

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

*4.3.1 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 [8, 22, 54]—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. 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 seriously limiting the sounds they recorded, most samples were recorded around the home.

*4.3.2 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 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.,* [33, 58, 78]). 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, that's going to be a different quality."* This motivated P2's decision 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.

*4.3.3 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 to be 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.'"*

*4.3.4 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.

## 4.4 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.

*4.4.1 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)[2] 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.[3] 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 that was not possible for 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]?"*

[2]The waveform displays the amplitude of sound rather than the pitch or frequency.
[3]Sample playback by the hearing first author revealed the sound of heavy rainfall at a similar volume to the thunder.

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."*

*4.4.2 Subjective Opinion of Sample Quality.* When reflecting on our characteristics for determining the samples' quality (Section 3.2.1), 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 to 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."*

*4.4.3 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 [15]. P10 suggested this issue caused to 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 chose to interpret the contents of their samples in comparison to others in their training set. 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). Review was also challenging for P7, and he saw potential for others to help: *"Relying on hearing people to feed the sound to a machine, [...] that might be better."*

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

## 4.5 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 [34] (Figure 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 used the *spectrogram* visualization briefly 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.*[4] *[...] 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."*[5]

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 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 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 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* (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 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 was in favor of automatic assessment, saying simply, *"I don't trust myself when it comes to the sound."* All participants selected post hoc *background noise feedback* (3d) and 11 added *overlapping sound feedback* (3e), hoping both could alleviate *"doubt"* (P1) over the samples' contents and with *"determining whether to re-record"* (P4).

[4] An audiogram displays the results of a pure-tone hearing test, the gold standard measure of hearing loss [75] 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/

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

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

*4.5.1 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 in comparison to a larger training set.

## 5 DISCUSSION

This study confirms the potential for non-expert DHH users to train personalizable sound recognizers (as identified in past work [58]) 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 real-world sounds, and (3) sense-making strategies that they use to interpret audio data in this context. Here, we discuss implications of our findings, opportunities for future work, and the limitations of our paper.

### 5.1 Technical Implementation

Our work fits within a supervised ML context in which audio samples are captured and labeled by DHH end users 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.,* [11, 48]) and specifically for DHH users [37, 38, 67] adapt deep learning approaches from computer vision—such as *VGG* [68] or *ResNet* [31]—and use transfer learning [72] to train on a large dataset of sounds such as AudioSet [26], FreeSound [25], or field recordings. For example, *SoundWatch* [38] 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 [51]. An initial personalizing step might be to adapt such a model to enable fine-tuning [79] 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 [58]. Meta-learning [24] 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 that are trained from an end-user's samples, our results suggest that systems intended for DHH users should allow for other data sources as well. For example, participants recorded few samples of urgent sounds during our study (Figure 4) despite this category being the most widely requested for sound awareness tools by DHH users [8, 22]. 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 [58] design that allows the choice between a recording tool or AudioSet [26] search.

DHH users who desire a personalized model but feel unqualified to record samples themselves, 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 [44]. 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 [60]. 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.,* [38]). 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 [35]); 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 use of 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.,* [20]) 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) [19], and future work intending to leverage the benefits of human-in-the-loop systems for DHH users should explore alternatives.

## 5.2 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. In this section, we propose possible solutions to these challenges through specialized instruction, enhanced audio visualization, and additional feedback to aid in 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 [45, 73], and instructional scaffolding can improve their understanding and satisfaction with personalized ML tools [46]. Prior work on non-expert ML use often provides scaffolding guidelines; for example, Yang *et al.* [78] 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 also 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 [32]) 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 an 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 [8], and most of our participants agreed the Rev app's [63] waveform visualization was a crucial for recording in situ. However, breakdowns in our participants' waveform sense-making highlights the potential for visualizations that are more intuitive 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 [12, 35] and DHH [54] novices can have with spectrograms. Yet these visualizations are shown to be powerful for experienced users [71]—including DHH ones, such as P14—and many sound recognizers extract features directly from spectrograms [32]. 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 [14], or explore new visualizations based on audiograms (2D frequency-volume graphs that are widely used in hearing loss testing [75] 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 are unable to 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 [35, 59] 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 [52] may separate sources or negate the impact of ambient sound [17]—additional feedback that informs DHH users about these artifacts may greatly enhance a DHH user's insight into the contents of the sample.

### 5.3 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 is reflective of 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 [7]. However, our study reiterates prior work showing the strong situational value that a sound recognizer can provide for some DHH people [8, 37, 38, 67], 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 also preserving a user's cultural preferences. In addition, although we designed our study with the belief 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 both hearing and DHH family and friends when recording which, in combination with collaborative benefits seen in a workshop setting [58], suggests that *interdependent* usage may be natural to some DHH users.

### 5.4 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.* [58] 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 people with high-speed internet access and enough time for 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 a variety of locations and allow users to solicit feedback from hearing people.

## 6 CONCLUSION

In this paper, we present an empirical account of the experience people who are d/Deaf and hard of hearing (DHH) may record and interpret sound samples for training a personalizable sound recognizer. Our findings demonstrate the need for specialized instructions to fill gaps audiological expertise, carefully-selected audio visualization that aligns with user intuition, and clear feedback to reveal sample diversity. In addition, our paper highlights the potential for several types of human-in-the-loop sound recognition systems with DHH users to be explored in future work. Our work has implications for sound designers, accessible technology researchers, and machine learning developers.

## ACKNOWLEDGMENTS

We thank Raja Kushalnagar for support with participant recruitment. This work was supported by the National Science Foundation under Grant No. IIS-1763199, the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1762114, and by University of Washington CREATE.

## 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*. Association for Computing Machinery, New York, NY, USA, 337–346. https://doi.org/10.1145/2702123.2702509

[5] Apple. 2020. iOS 14 - Features - Apple. Retrieved September 15, 2020 from https://www.apple.com/ios/ios-14/features/

[6] Audacity Team. 2020. Audacity(R): Free Audio Editor and Recorder. Retrieved July 19, 2020 from https://audacityteam.org/

[7] 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.

[8] 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

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] 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

[11] 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

[12] 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

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

[14] 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 International Conference on Computational Intelligence and Data Science.

[15] 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

[16] 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

[17] 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.

[18] 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

[19] 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 (jul 2018), 1–37. https://doi.org/10.1145/3185517

[20] 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

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

[22] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300276

[23] Leah Findlater, Steven Goodman, Yuhang Zhao, Shiri Azenkot, and Margot Hanley. 2020. Fairness Issues in AI Systems That Augment Sensory Abilities. *SIGACCESS Access. Comput.* 125, Article 8 (March 2020), 1 pages. https://doi.org/10.1145/3386296.3386304

[24] 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.

[25] 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.

[26] 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

[27] 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

[28] Google. 2020. Audio Model - Teachable Machines. Retrieved July 19, 2020 from https://teachablemachine.withgoogle.com/train/audio

[29] Google. 2020. Important household sounds become more accessible. Retrieved October 12, 2020 from https://blog.google/products/android/new-sound-notifications-on-android/

[30] Sébastien Gulluni, Slim Essid, Olivier Buisson, and Gaël Richard. 2011. An Interactive System for Electro-Acoustic Music Analysis. In *Proc. ISMIR*. 145–150.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[32] 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.

[33] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376428

[34] 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

[35] 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* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 67–77. https://doi.org/10.1145/3377325.3377483

[36] 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

[37] 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

[38] 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 *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 30, 13 pages. https://doi.org/10.1145/3373625.3416991

[39] 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

[40] Hernisa Kacorri. 2017. Teachable Machines for Accessibility. *SIGACCESS Access. Comput.* 119 (nov 2017), 10–18. https://doi.org/10.1145/3167902.3167904

[41] 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

[42] 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

[43] 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 (jul 2018), 1–23. https://doi.org/10.1145/3214366

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

[45] 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

[46] 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

[47] 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

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

[49] 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)*. Association for Computing Machinery, New York, NY, USA, 83–95. https://doi.org/10.1145/3308561.3353799

[50] 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

[51] Makeability Lab. 2020. SoundWatch. Retrieved November 8, 2020 from https://github.com/makeabilitylab/SoundWatch

[52] 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.

[53] 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

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

[55] 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

[56] 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

[57] 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.

[58] 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* (Tallinn, Estonia) *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 82, 12 pages. https://doi.org/10.1145/3419249.3420157

[59] 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. arXiv:1802.05316 [cs.HC]

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

[61] 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

[62] 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

[63] Rev.com. 2020. Voice Recorder App | Audio Recording App. Retrieved July 19, 2020 from https://www.rev.com/voicerecorder

[64] James Robert, Marc Webbie, et al. 2018. Pydub. http://pydub.com/

[65] Prem Seetharaman, Gautham Mysore, Bryan Pardo, Paris Smaragdis, and Celso Gomes. 2019. VoiceAssist: Guiding Users to High-Quality Voice Recordings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290605.3300539

[66] 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

[67] 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 (jun 2017), 17:1—-17:21. https://doi.org/10.1145/3090082

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

[69] 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

[70] 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

[71] 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 (2009), 42–50. https://doi.org/10.1111/j.1557-9263.2009.00204.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1557-9263.2009.00204.x

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

[73] 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

[74] 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

[75] 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.

[76] 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

[77] 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

[78] 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

[79] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792* (2014).

[80] Zoom Video Communications. 2020. Video Conferencing, Web Conferencing, Webinars, Screen Sharing. Retrieved July 19, 2020 from https://zoom.us