SPECTRA: Personalizable Sound Recognition for Deaf and Hard of Hearing Users through Interactive Machine Learning

Steven M. Goodman
Human Centered Design and Engineering
University of Washington
Seattle, Washington, USA
smgoodmn@uw.edu

Emma J McDonnell
Human Centered Design and Engineering
University of Washington
Seattle, Washington, USA
ejm249@uw.edu

Jon E. Froehlich
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, Washington, USA
jonf@cs.uw.edu

Leah Findlater
Human Centered Design and Engineering
University of Washington
Seattle, Washington, USA
leahkf@uw.edu

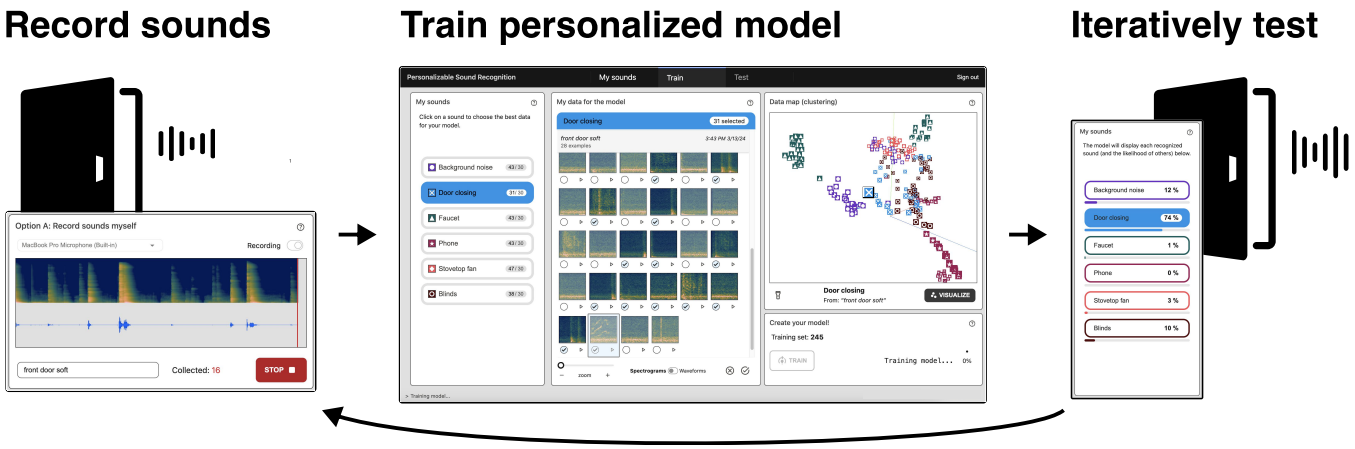

Figure 1: Overview of the SPECTRA pipeline. In an interactive machine learning training workflow, users collect audio data samples (left), filter their data into a training dataset (center), and assess their model's performance in a live environment (right). The design includes key elements to support the needs of DHH users during this process, including spectrogram and waveform audio visualizations of audio, data annotating to save useful contextual information, and an interactive clustering visualization of their dataset.

Abstract

We introduce SPECTRA, a novel pipeline for personalizable sound recognition designed to understand DHH users' needs when collecting audio data, creating a training dataset, and reasoning about the quality of a model. To evaluate the prototype, we recruited 12 DHH participants who trained personalized models for their homes. We investigated waveforms, spectrograms, interactive clustering, and data annotating to support DHH users throughout this workflow, and we explored the impact of a hands-on training session on their experience and attitudes toward sound recognition tools. Our findings reveal the potential for clustering visualizations and waveforms to enrich users' understanding of audio data and refinement of training datasets, along with data annotations to promote varied data collection. We provide insights into DHH users' experiences and perspectives on personalizing a sound recognition pipeline. Finally, we share design considerations for future interactive systems to support this population.

CCS Concepts

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

Keywords

Deaf and hard of hearing, sound recognition, accessibility, interactive machine learning

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713294

ACM Reference Format:
Steven M. Goodman, Emma J McDonnell, Jon E. Froehlich, and Leah Findlater. 2025. SPECTRA: Personalizable Sound Recognition for Deaf and Hard of Hearing Users through Interactive Machine Learning. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3713294

1 Introduction

Sound carries rich information about events around us, but it may go unnoticed or be inaccessible to individuals who are Deaf, deaf, or hard of hearing (DHH). Prior work shows that many DHH people desire sound recognition technologies to support personal safety (*e.g.,* footsteps), social engagement (*e.g.,* nearby voices), and everyday tasks (*e.g.,* monitoring home appliances) [4, 15, 30]. To meet this need, sound recognition tools have proliferated, both in the research literature (*e.g.,* [31, 46, 57]) and in commercial applications—for example, Android and iOS smartphones support sound recognition for common sounds like doorbells, running water, and dog barks.

Despite these advances, DHH users have expressed a need for improved sound recognition accuracy and support for a wider range of sound categories [28]. One challenge is that the value of sound information is highly contextual: hearing identity [15], social context [4], physical location [19], and individual preferences [15] can all influence how a DHH user may benefit from sound information. These contextual differences could be addressed by empowering DHH individuals to personalize a sound recognition model themselves [4, 20] or fine-tune a model to their local soundscape [30]. Personalization also has the potential to improve accuracy by providing examples specific to an individual user's context, such as the sound of *their* appliance or alarm rather than a generic one.

However, an open question lies in how to effectively support a DHH user—who does not have full access to a sound themselves—in capturing and selecting suitable audio data to train a machine learning (ML) model [20]. Android and iOS recently introduced the ability to add custom sound categories [21] or tune the model for specific alarms and appliances [2], respectively, through a brief recording process. While this approach is simple, allowing users with or without ML expertise to engage more directly with the machine learning pipeline through an interactive machine learning (IML) approach provides a sense of transparency and control [12], which can positively impact trust, satisfaction, and long-term use [1, 38]. Prior work has begun to investigate the potential of IML for sound recognition systems for DHH users, but, in one case, did not provide non-auditory feedback to DHH participants [50] and in a second, only focused on data collection within the ML pipeline [20]—not including model training, evaluation, and iteration.

In this paper, we introduce and explore *SPECTRA—Sound Processing and Enhanced Custom Training for Recognition Assistance*—an interactive pipeline for the accessible creation of personalized sound recognition models. Our human-centered approach merges IML design guidelines [13, 52, 67] and the needs of DHH users [20, 50] to train a personalized sound recognizer via IML. To evaluate how SPECTRA supports DHH users in engaging in IML, we recruited 12 DHH participants who each trained a personalized sound recognition model for their home soundscape. We examined how spectrogram and waveform visualizations, interactive clustering, and rich text annotations can support DHH users across an interactive training cycle, and how their experience with these mechanisms can shape performance expectations, technical understanding, and confidence. We also investigated the impact of the experience on participants' perceptions of and attitudes toward personalizable sound recognition models.

Our findings reveal new insights to support DHH users in personalizing sound recognition models, including demonstrating how interactive data clustering, in combination with waveforms, can enhance DHH users' understanding of audio data, identification of outliers, and refinement of training datasets. We show the value of non-auditory data representations for DHH users at different stages of an end-to-end training cycle (data collection, training data selection, model testing) and explain how they incorporate this information into their reasoning about sound models. Our results also reaffirm prior work showing DHH users' preference for waveforms when recording [20]—while expanding on their value when selecting training data and testing—and show how users' training strategies develop through use [50]. Finally, we provide insights into DHH users' experiences and perspectives on personalizing a sound recognition model.

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

2 Related Work

We situate our research within prior work on sound awareness tools, human-centered ML, and cultural and contextual factors that influence sound awareness.

2.1 Sound Awareness Tools and Technologies

Prior work in sound awareness has introduced systems for automatic captioning [27, 51], sound alerting [34, 37], and vibrational feedback [42, 60], including smartphone [43, 47], IoT [29], and wearable form factors [26, 68]. Surveys of DHH individuals reveal a widespread desire for sound recognition tools that can notify when a sound is detected [4, 15, 28, 31]. Following advancements in digital signal processing and machine learning, recent work has aimed to provide broad sound recognition support by employing pre-trained classification models [30, 31, 46, 57]. Jain *et al.* [30] installed a tablet-based system for recognizing 19 sounds in four homes, observing concerns over inconsistent classification and desires to personalize the system for sounds specific to each home. We build on this work, focusing on in-home personalization.

DHH users frequently request personalization options for as-needed support of their individual needs (*e.g.,* [15, 44, 47]). Some pre-trained sound recognition models have allowed DHH users to filter notifications for certain sounds [26, 31]—a customization option that is now available on Apple and Android smartphone platforms. Going a step further, Bragg *et al.* [4] conducted a Wizard-of-Oz usability study of a smartphone app to train a custom model

from end-users' recordings; DHH participants responded positively to the app's workflow, but the experience did not include a functional model. Continuing the thread, Jain *et al.* [28] surveyed 472 DHH Android users about the platform's sound recognition feature, confirming users' desire for personalization options. The authors developed ProtoSound, a training pipeline with technical considerations for DHH users (*e.g.*, limited data, contextual flexibility)—but they did not evaluate it with DHH users [28]. So far, researchers have not tested a functional, end-to-end system that meets the needs of DHH users.

2.2 Human-Centered Machine Learning

Human-centered machine learning aims to design and build automated systems that can fulfill user goals, fit user-specific contexts, and accommodate people without programming experience [14, 53]. Several approaches have emerged to enable non-experts to build their own ML models. *Automated Machine Learning* (AutoML) (*e.g.*, [10, 64]) systems allow novice end-users to provide a large batch of labeled data, while traditional ML tasks—such as feature engineering and model selection—are completed automatically [12, 66]. In contrast to AutoML's black box approach, interactive machine learning (IML) treats end-users as "humans-in-the-loop" who iteratively engage in building and refining ML models [1, 13, 52, 59]. An IML workflow (like SPECTRA's) involves a quick loop between model training, feedback, and usage, where the user may provide indicative samples, describe salient features, or select high-level model parameters [13, 59]. Interactive machine *teaching* [52, 69] takes IML engagement one step further by positioning the human-in-the-loop in the role of the model's teacher, emphasizing human expertise to guide machine learning [63]. While each paradigm can help to guide the design of personalizable sound recognition tools, they also assume that an end-user has domain expertise and can readily interpret their model's underlying data—assumptions that may not hold for DHH users and audio data.

In the field of accessibility, human-centered ML applications hold the potential for disabled users to personalize data-driven assistive technology that meets their individual needs [32]. 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 [16, 49]. 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.*, [33, 58]) first examined and more recently began addressing through active feedback techniques to assist in the image capture step [24, 41].

Human-centered ML work with DHH users includes a workshop study by Nakao *et al.* [50], which sought to characterize ML understanding among DHH participants through their collaborative use of a sound recognition interface. Participants were uncertain about the contents and quality of sound data due to the absence of non-auditory feedback (*e.g.*, visualizations) within the system. Goodman *et al.* [20] explored DHH participants' experience capturing real-world sound data with a waveform-based recording app; while the visualization assisted with data capture, users expressed uncertainty about overall dataset quality and nuanced sound categories. In light of these challenges with data interpretation, our work explores how non-auditory data representations can better support DHH users across an end-to-end training cycle.

General IML research for audio has primarily focused on sample annotation and labeling (*e.g.*, [25, 35, 36, 56]). For interactive sound recognition, Ishibashi *et al.* [25] explored visualization options (*e.g.*, spectrograms, thumbnails) for browsing large sets of unlabeled audio samples via a clustering interface. Google's Teachable Machine [7] allows non-expert users to quickly train a personal sound recognition model with their own audio samples, but it provides limited audio visualization (low-resolution spectrograms) and lacks information on the quality of a user's training set (*e.g.*, clustering feedback). Nakao *et al.*'s work [50] explored a comparable workflow (without visualizations) with non-expert DHH users, allowing them to create training sets by recording or selecting from a sound library. After a shared hands-on experience, DHH participants identified additional use cases and showed an improved understanding of ML; however, some found it challenging to review samples and define classes for sounds they were familiar with but unable to hear.

This prior work—in combination with others [4, 28]—begins to outline an interface design space for personalizable sound recognizers for non-expert DHH users. As a next step, we built and evaluated a specialized IML workflow tailored to the unique needs of DHH users, advancing understanding of how non-auditory data representations can support this population during interactive sound recognition tasks and yielding insights towards the design of future tools in this area.

2.3 Cultural and Contextual Factors of Sound Awareness

Designing effective sound awareness technology requires understanding the DHH population's wide-ranging preferences. An individual with hearing loss may identify as Deaf (capital 'D'), deaf, or hard of hearing [39, 48]. Individuals who identify as Deaf follow an established set of norms, behaviors, and language [40], which contrasts with hard of hearing or deaf individuals, for whom deafness is primarily an audiological experience. While prior work has shown widespread interest in sound awareness among DHH people [4, 15, 29, 44], this interest is modulated by cultural factors. An online survey by Findlater *et al.* [16] with 201 DHH participants found that people who prefer oral communication are more interested in sound awareness than those who prefer sign language.

While accounting for the diverse perspectives of DHH people, prior work also highlights several general preferences among DHH users. When discussing sounds of interest, DHH users generally rank urgent sounds (safety-related alarms, sirens) as most important, followed by those indicating others' presence (door knocks, footsteps) and appliance alerts (oven timers, pop-up toasters) [4, 15, 44, 57]. As previously mentioned, however, the relevance of certain sounds can depend on one's social context [4, 15, 26] and physical location [19]. Overall, the most desired dimension of sound information is identity (*i.e.,* what sound is occurring), which users prioritize compared to other characteristics like volume or duration [4, 15, 19]. However, context may again influence preferences; for example, sound identity may be adequate in the home [30], but directional indicators hold more importance while mobile [46].

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

To investigate how visualization techniques can support DHH users in personalizing their own sound recognition models, we built SPECTRA (Sound Processing and Enhanced Custom Training for Recognition Assistance), a prototype IML web application. The pipeline's design was informed by related sound awareness literature [4, 20, 28, 50] and guidelines for human-centered ML systems [13, 52, 55, 67]. SPECTRA has a three-step workflow: users first generate a training dataset, then edit the training set and generate a model, before testing the model's real-time sound recognition capacity. In this section, we describe the implementation of each step and outline the pipeline workflow and functionality.

3.1 Spectrogram and Waveform Feedback

SPECTRA uses high-fidelity waveform and spectrogram visualizations to convey audio data to DHH users (Figure 2). Waveforms show the amplitude—or loudness—of audio over time and are common in audio recording, editing, and playback software. DHH participants in prior work reported waveforms were intuitive and useful for capturing audio examples, though the amplitudinal feedback alone was inadequate for verifying the recordings' quality (*e.g.*, no co-occurring sounds) [4, 20]. They further requested that the visualizations remain active before and after recording to monitor the ambient soundscape, while audio playback helped those with residual hearing to analyze waveforms with unclear meanings [20]—both of which are included in SPECTRA.

Spectrogram visualizations offer greater information throughput by visualizing both amplitude and frequency and are commonly used for discriminating noises in environmental soundscapes (*e.g.,* [11]). While spectrograms can be powerful data interpretation tools for experienced users [8], DHH participants had a mixed response following brief use in a lab setting [20]. Models trained with SPECTRA—like many sound recognition models (*e.g.*, [28, 57])—take Mel spectral features (*i.e.*, frequencies bucketed to approximate human hearing) as input, meaning that users viewing spectrograms see the same audio properties the model uses to make decisions.[1] We include both visualizations to cater to DHH individuals' diverse preferences and learning styles and explore their effect on users' decision-making about a training dataset.

3.2 Interactive Data Clustering

SPECTRA includes a three-dimensional data clustering visualization to help DHH users understand and refine an audio dataset (Figure 3). We draw from prior work on interactive data clustering [3], including sound clustering with hearing users [25], to address the unique challenges DHH users face in an iterative training process. Goodman *et al.* found DHH individuals had issues with discerning variations in sounds (*e.g.*, porcelain *vs.* metal faucets) and anticipating how audible differences will affect model performance [20].[2] SPECTRA's clustering visualization complements the waveform and spectrogram displays—which show individual audio instances—by rendering the structure and diversity of the broader dataset. We employ UMAP dimensionality reduction [45] to project high-dimensional spectral audio features into an embedding space, where similar examples colocate while distinct examples separate.

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

3.3 Rich Data Annotations

During data collection, SPECTRA allows users to annotate their recordings with textual descriptions, capturing contextual details (*e.g.*, water running from bathroom *vs.* kitchen sink). Annotations serve as a form of semantic metadata for users, separate from model labels. Many real-world sounds vary depending on their source, production method, or environment—a challenge when personalizing sound recognition models, where users need to provide a representative dataset for the model to generalize to their environment. DHH users in prior work questioned the meaning of differences among their recordings and the impact of sound variations on their models' performance [20]. SPECTRA encourages users to identify and capture variations of each sound (*e.g.*, faucet → stream, drip), drawing from the concept decomposition process of the interactive machine teaching paradigm [52, 63]. Annotations allow DHH users to document domain expertise not readily apparent with SPECTRA's visual feedback alone, such as clarifying subtle differences in waveforms/spectrograms or supporting reasoning about loose or separated clusters with the same sound label. Annotations are displayed alongside SPECTRA's visualizations during data selection, serving as a memory aid and reasoning tool to make sound data more understandable to DHH users.

3.4 SPECTRA Implementation and Workflow

We built SPECTRA using *Node*[3] and *Svelte*[4] from a fork of *Marcelle.js*[5], an open-source toolkit for creating ML workflows and interfaces [17]. To enable transfer learning from a pre-trained sound

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

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

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

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

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

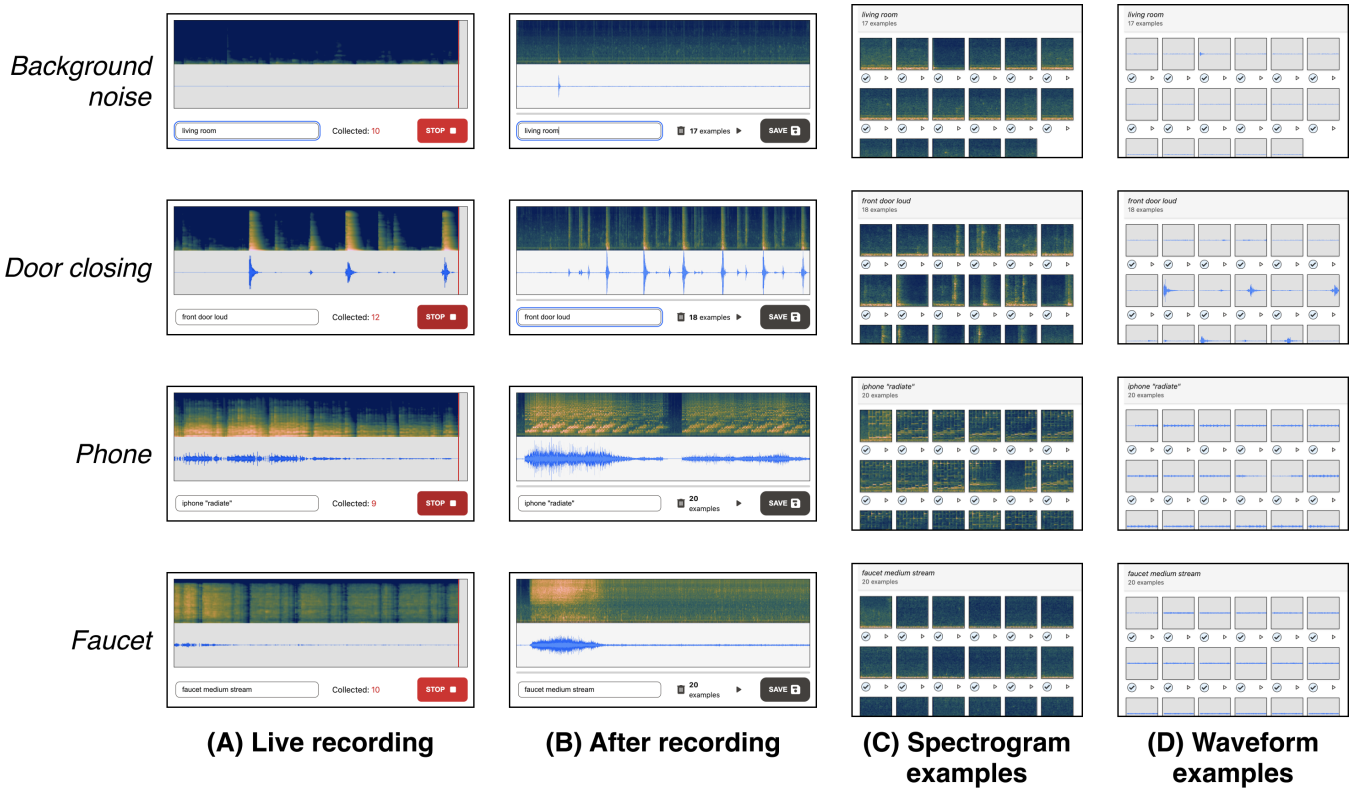

Figure 2: Spectrogram and waveform visualizations for four different sounds within the SPECTRA workflow. (A) While recording, a stacked spectrogram and waveform display streaming audio input over a 10-second window. (B) After recording, the stacked visualization shows the full duration of the captured audio. When selecting training data, each recording is segmented at 1-second intervals, which users can choose to display as (C) spectrogram or (D) waveform icons.

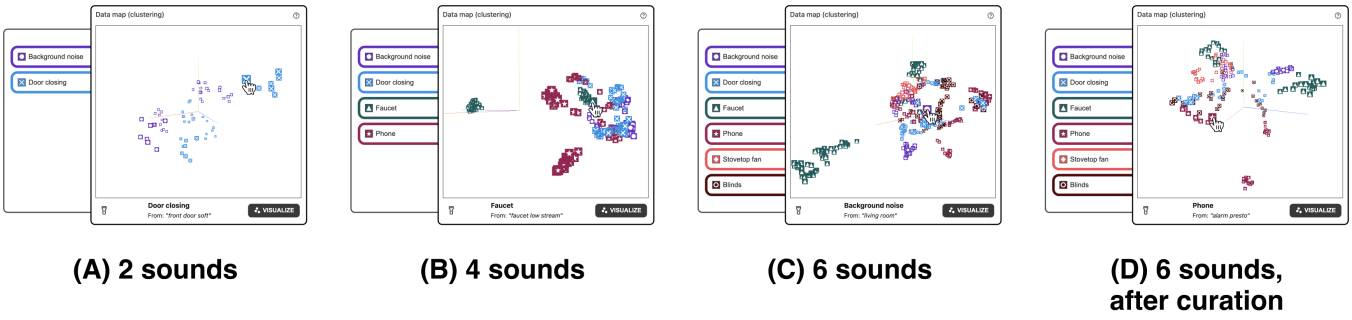

Figure 3: Clustering visualizations within SPECTRA, showing the structure of unedited audio datasets with (A) two sounds, (B) four sounds, and (C) six sounds. As the size/variety of the dataset increases, so does visual complexity. (D) Curating the raw six-sound dataset (*i.e.*, removing mislabeled examples) shows greater separation for many clusters, though some overlap persists. Note: Hovering over a data point displays the example's label and the annotation (metadata) of its parent recording.

model, SPECTRA is powered by the *Speech Commands API*[6] from *Tensorflow.js* [62], which employs a convolutional neural network (CNN) pre-trained on the Speech Commands dataset (50K examples from 20 classes) [65]. CNNs are commonly used in sound recognition due to their ability to learn complex patterns in audio data [22]. SPECTRA uses the API's transfer learning functionality—where a pre-trained model is re-used as a feature extractor for new classes, reducing training time and resources—to apply the speech-trained base model to the environmental sound domain. We chose this

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

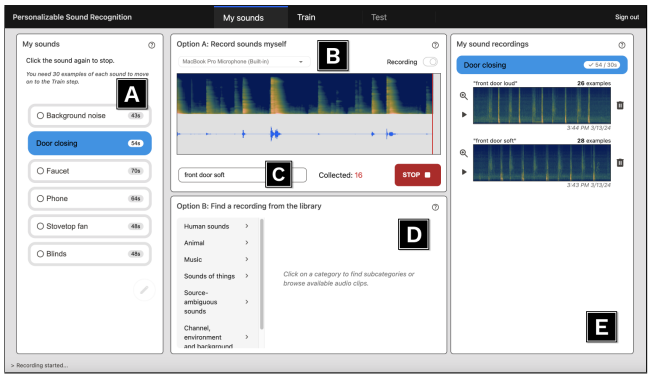

Figure 4: The "My Sounds" page, used for data collection. (A) Users select a sound class to begin collection. (B) They can record with a live waveform and spectrogram of their device's microphone input, and (C) annotate the recording with relevant contextual details. (D) Alternatively, users can select pre-recorded examples from a categorized library of video clips. (E) Users can review collected recordings for the selected class.

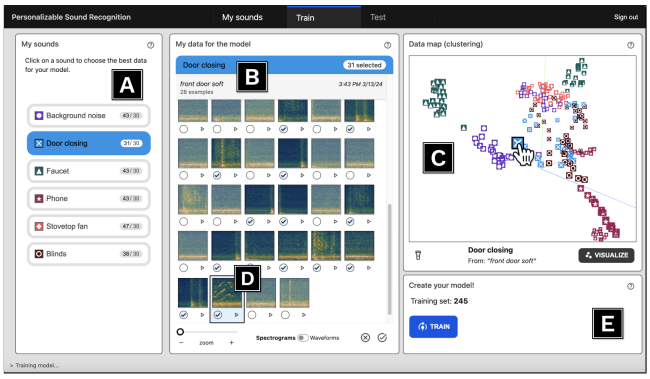

Figure 5: The "Train" page, used for training data curation. (A) Users select a sound class to filter their sampling set. (B) Examples are shown as 1-second spectrograms, which can be toggled for inclusion in the training dataset. (C) Users generate a three-dimensional clustering of the selected training data and can rotate or zoom to inspect the clusterings. Hovering on a point will show its label and text annotation; (D) clicking on it will highlight the example in the data selection panel. (E) Users train a model with the selected training dataset.

library for its ease of development, rapid prototyping capabilities, suitability for in-browser use, and lightweight package. These allowed us to focus on the interactive aspects of the system and study how DHH users engage with the IML workflow.

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

While prior work has explored mobile devices for recording audio for personalizable sound recognition [4, 20], we focus on the entire IML workflow and thus designed SPECTRA for laptop/desktop screens. We do not assume this is an ideal format for end-users; instead, we leverage the large screen size to present multiple high-fidelity visualizations in tandem (waveform, spectrogram, and/or clustering) and learn about salient information to assist DHH users when personalizing a sound recognizer. SPECTRA's UI is organized into three tabs ("pages") corresponding to different stages of the IML workflow [13]: (1) planning and data collection; (2) data curation and model training; and (3) iterative model testing.

3.4.1 Planning and Data Collection. SPECTRA users start at the "My sounds" page (Figure 4), which aligns with the planning and data collection stages of a typical interactive ML workflow (*e.g.*, [13]). Users first define and create placeholder classes for desired recognizable sounds in the "My sounds" panel (*e.g.*, "my dog barking" or "stovetop fan"). SPECTRA currently supports adding up to 10 distinct classes of sounds. The "My sounds" panel (Figure 4a) is available across all three pages on the left sidebar. The user then selects a specific sound class to initiate data collection, which activates the center and right-side UI panels.

To collect data, users navigate to the center panel (Figure 4b) and click the "Start listening" toggle. Live microphone data is then visualized via the waveform and spectrogram visualizations (but not yet recorded). After resolving any unwanted background noise, the user can collect data with the "Record" and "Stop" buttons. While recording, SPECTRA shows users a running count of the collected examples (*i.e.*, number of 1-second spectrogram increments). Users can add text annotations before or after recording to note additional information, such as sound variations (*e.g.*, bathroom *vs.* kitchen sink) or unplanned sound activity (*e.g.*, a cat meowing mid-recording) (4c). Alternatively, for hard-to-record or unavailable sounds, (*e.g.*, sirens), users can import existing recordings from the *AudioSet* library of categorized YouTube clips [18] (Figure 4d). If they are satisfied with the recording, users can save it to their sampling set, which shows up on the right-side pane under "My sound recordings" (Figure 4e). Before users can move on to SPECTRA's "Train" page, they must collect at least 30 examples of each sound (encouraged to be varied across a few 5-10 second recordings)—a threshold selected to ensure sufficient data for model training without overburdening users.

3.4.2 Data Curation and Model Training. In the second stage, users navigate to the "Train" page, where they review their data, refine the training dataset, and train a model (Figure 5). On this page, when users select a sound class from the "My sounds" panel (Figure 5a), the center panel populates with one-second examples of that sound (Figure 5b). The examples, grouped by recording, are presented as spectrograms, with the option to switch to waveform

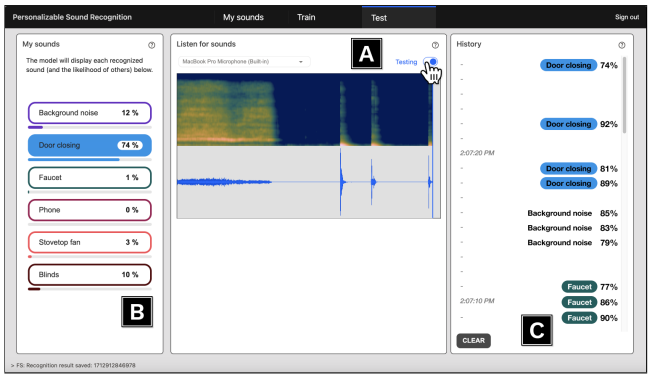

Figure 6: The "Test" page, used for practical assessment. (A) Users toggle their microphone "on" to begin streaming recognition. (B) Users produce sounds in the environment and observe the model's predictions as confidence scores (with a bar graph and percentage). (C) A vertical timeline also shows the highest-scored sound at each second of the previous two minutes.

visualizations. Users can select which examples to include or exclude in the training dataset, with all examples included by default.

SPECTRA's emphasis on data iteration aligns with practices observed among expert ML practitioners, who typically prioritize dataset refinement over changes to the models themselves [23]. To support users' understanding of their dataset, SPECTRA includes an interactive data clustering panel ("Data Map"; Figure 5c). As they filter out low-quality or unrepresentative examples, users can generate new embeddings of the updated dataset in latent space to monitor how the overall structure changes. The clustering visualization uses UMAP for dimensionality reduction [45] (enabled by *UMAP-js*[7]) of the high-dimensional Mel spectrograms as training features. For simplicity, we chose pre-set UMAP parameters[8] after testing with several audio datasets. These 43×232 arrays are reduced to a 1×3 size and plotted in 3D space with *ScatterGL*[9], using a unique symbol to represent each sound class in the embedding space (located next to that class in the "My Sounds" panel). While embeddings generated from SPECTRA's base model's feature space may provide better scalability and align more with perceptual similarity, our dimension-reduction of the Mel spectrograms provided a fast and light-weight method that was acceptable for the constraints of our evaluation—serving as a design probe into how to support DHH users' reasoning about algorithmic interpretations of sound.

Users explore the clustering visualization by rotating (click and drag) or zooming (scroll), and they can select an individual point to highlight its corresponding example in the selection panel for further inspection (5d). Users can iteratively adjust the training set and observe its effect on clustering, moving toward more clearly differentiated sound classes. Once satisfied with the refined training set, clicking the "Train" button creates a new model, which users assess on the third page.

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

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

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

3.4.3 Model Testing. In the third and final stage, users navigate to the "Test" page, where they can assess the practical performance of their personalized model's latest iteration (Figure 6). Users activate a toggle switch to initiate streaming recognition (Figure 5a) and can then produce sounds to evaluate the model's predictions in a live environment along with both waveform and spectrogram visualizations. This practical assessment allows users to evaluate the model's performance in its intended use context and under realistic conditions. In the "My sounds" panel on this page, real-time predictions are displayed with a confidence score and bar chart below each class label (5b). Sounds recognized with a confidence score of 70% and above are added to a history panel on the right with a timestamp (5c).[10] After testing, users may return to the previous pages to add or remove sound labels, collect additional sampling data, or modify their training data and train a new model.

4 Evaluation

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

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

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

4.1 Participants

We recruited 12 DHH participants from two university-maintained study recruitment email lists and snowball sampling—see Table 1 for demographic details. Five participants identified as Deaf, five as hard of hearing, and two as deaf. Four participants reported using hearing aids; four used cochlear implants; two used both. We included two technology-related screening requirements: weekly laptop or desktop computer use and daily smartphone use for tasks other than phone calls and text messaging. Participants also needed access to a laptop or desktop computer at home with a working camera and microphone, a stable Internet connection for videoconferencing, and Google Chrome to use SPECTRA. Nine participants used a laptop; three used a desktop. Participants received an $80 gift card as compensation for the 120-minute session.

We did not screen for prior ML experience among participants, as we designed SPECTRA to assist all DHH users interested in personalizing a sound recognition model. As a whole, participants self-reported moderate confidence in explaining basic principles

[10]The 70% threshold for sounds added to the history panel reflects a need for higher confidence in our study's 6-class models when compared to lower confidence thresholds used for larger-class models in prior work (*e.g.*, 50% [31], 60% [28]).

Table 1: Demographics of study participants. HH = hard of hearing.

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

of machine learning (*e.g.*, training data, predictions, models)—on a 7-point scale, the average response was 4.8 (*SD*=1.3, range=3-6). Five participants mentioned hands-on ML experience during the session (Table 1)—ranging from a brief university project (P6) to regular ML use for research (P8)—but none had worked on audio models before our study. Our study explored changes in their perceptions of audio models in particular (*e.g.*, tolerable performance levels before/after use), and we note users' experience where relevant in our findings.

4.2 Study Procedure (120 min)

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

4.2.1 Tutorial and Pre-Use Interview (30 min). Sessions began with five minutes for Zoom setup and orientation, followed by a tutorial slide deck[11], which participants could navigate at their own pace. The tutorial, informed by prior work on IML systems for non-expert users [55, 67], provided an overview of sound recognition models' learning and decision-making, the differences between generalized and personalized models, and possible advantages of personalization. It then introduced each stage of the SPECTRA pipeline, with accompanying screenshots demonstrating recording, training data selection, model training, and assessment as well as explanations of spectrogram, waveform, and data clustering visualizations (*e.g.*, *"Examples closer together are more similar, while those further apart are more different"*). We encouraged participants to ask questions throughout the tutorial.

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

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

4.2.2 System Use (60 min). After completing the tutorial, participants accessed SPECTRA via a shared link and began screen-sharing with videoconferencing. We instructed participants to "think aloud" throughout system use and to freely voice any questions, observations, suggestions, or concerns that arose. For participants using ASL, comments followed or preceded system interactions, rather than occurring concurrently. To prevent significant barriers to training a model, the researcher provided troubleshooting and clarification support where appropriate; we noted these areas of friction and included them in our analysis.

We asked participants to train a model with five sound classes, plus "background noise" to provide a baseline for the ambient soundscape—a decision informed by Jain *et al.*'s survey in which a majority of Android sound recognition users desired notifications for six sounds or fewer in a single location [28]. To orient participants, we asked them to turn on their microphone and observe the visualizations while no one spoke or acted and then to see how they changed when intentionally making noise. We defined the soundscape in the absence of intentional noise as "background noise" for this study. We pre-selected three sounds (door closing, faucet, and phone) based on how easily they could be produced and the range of possible variations. Participants brainstormed and then selected the remaining two sounds.

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

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

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

4.2.3 Semi-Structured Interview and Rating Scales (30 min). We concluded the study with a post-use questionnaire and semi-structured interview. The questionnaire measured changes in confidence and performance expectations after use (mirroring statements on the pre-use ratings); satisfaction with recordings, training data, and model accuracy; and the usefulness of the text annotations, waveform, spectrogram, and clustering visualization. The interview explored overall satisfaction, experience with each step of the application, confidence in independent training, and opinions on the data exploration mechanisms and potential UI improvements.

4.3 Analysis and Positionality

We collected each participant's usage logs, audio recordings, and training datasets to further characterize their responses and experience with the pipeline. We used Zoom's automatic captioning to transcribe study data, relying on voiced interpretations as an accurate representation of signers' responses. We iteratively coded transcripts using reflexive thematic analysis [5, 6]. Our analysis was semantic and realist, and we developed themes using a mixed inductive and deductive approach; for example, we structured broader theme development around the distinct tasks at each step of personalizing a sound recognizer (*i.e.*, recording, choosing training data, testing), but we organically identified themes within each step. The first author read through the data, generated initial codes, and then applied these codes to data from two randomly selected participants. Another researcher reviewed the coded data and then met with the first author to discuss the codes and application strategy. The first author coded the remaining transcripts and generated themes from data excerpts collated from each code. A reflexive approach to thematic analysis emphasizes findings that are actively shaped by the research team's own social, cultural, and academic biases. All authors are hearing, while past collaborators—who contributed to the early system and study design—are Deaf or hard of hearing. All research team members have backgrounds in human-computer interaction, and many are computer scientists by training.

5 Findings

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

Table 2: Collected data and system usage for study participants. (Column A) Participants chose sounds to train in addition to the required "background noise", "door closing", "faucet", and "phone" sounds. (B) They captured at least one n-second recording of each sound. (C) Their recordings were segmented into 1-second examples to use as training data. (D) They selected a subset of the examples to train a model. (E) They generated clusterings to visualize different iterations of this subset.

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

5.1 Overview of Collected Data and SPECTRA Usage

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

For training and testing, participants spent an average of 20.7 mins (SD=5.7) on the "Train" page and 10.6 (SD=2.0) on the "Test" page. The training itself was interactive and iterative, with nearly half of the time spent focused within the clustering chart—on average, participants clicked on 7.5 clustering points and regenerated clusterings 4.5 times (SD=1.8, range=2-8) after making changes to their training dataset. Participants removed an average of 8.6 examples in their final training datasets (23% reduction), demonstrating the visualizations' influence on their decision-making. Overall, final training datasets averaged 37.7 examples per sound (slightly above the minimum requirement).

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

5.2 Pre-Use Perceptions and Expectations

In the pre-study questionnaire, most participants (N=7) expressed positive interest in automatic sound recognition technology ("likely" or "extremely likely" to use it), including for urgent (P7: *"smoke alarm"*), social (P1: *"someone arriving home"*), and appliance sounds (P9: *"oven timers"*)—aligning with prior work [4, 15]. Four participants remained "neutral", while P6 was "unlikely" to use such technology. Participants also identified several uses for personalized models, most commonly related to identifying specific speakers (N=5) and nuanced pet sounds (N=4).

Five participants reported using sound notification features on their smartphones, albeit infrequently (semi-monthly or less), citing limited sound selection support and inaccurate recognition as key issues—echoing past findings [28]. Only P5 had previously attempted to add a custom sound class (on the iPhone), but even here, he had experienced issues: *"It's like, 'Someone's knocking at the door,' but it's actually my roommate, cutting with a knife"*.

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

5.3 Using SPECTRA to Interactively Train a Personalized Sound Recognition Model

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

5.3.1 Waveform and Spectrogram Visualizations. Sound visualizations are essential for making audio data accessible to DHH users [20, 50]. However, how best to visualize sound to support interactive training of a sound recognition model is an open research question—especially for users who may have different mental models of sound and/or lack access to the auditory channel itself. Thus, drawing on prior work [20], we designed SPECTRA to use waveform and spectrogram visualizations for both streaming and static sound information when recording audio, selecting training examples, and testing a new sound recognition model.

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

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

5.3.2 Interactive Data Clustering. Participants deemed the data clustering visualization critical to the IML process—usefulness rating: *avg.*=6.3, *SD*=1.2—primarily because it helped provide transparency, feedback on audio recording quality, and assisted with iteratively refining the training dataset. As P7 stated, *"[It's] a good depiction of where everything lies and how the model is looking at it"* (P7). Participants saw clusters as useful when trying to understand the consistency of samples in a given class as well as distinctiveness across classes: *"This is the door closing, and it's clustered right here, so I know that I did a good job"* (P8). Clustering also served as a bridge for participants to begin considering their data in terms of how it may impact the behavior of a ML model – P2 iterated on their data set until the clustering seemed *'more clear [now]; it doesn't seem as if there's a lot of overlap"*. For P1, watching the clustering change after adjusting the training dataset felt affirming: *"[It] was satisfying to see, 'Okay, like it's actually working; what I'm doing'"*). In contrast to prior work where DHH participants expressed uncertainty about the quality of self-collected training data [20], using the clustering visualization increased participants' confidence that they had collected their desired data; upon a final review of his clusterings before training, P4 said, *"I feel pretty good about this—the [examples] that are remaining."*

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

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

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

seem like the [background noise] is now pulled apart a little bit, and the [faucet]").

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

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

5.4 Post-Study Questionnaire and Interview

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

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

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

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

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

5.4.5 Strategies to Improve Model Performance. Participants' experiences with SPECTRA shaped their perceptions of and intentions for future use of IML for personalized sound recognition. This engagement with the system caused them to grapple with the complexity of training an accurate sound recognition model—seven participants' self-reported understanding of how to improve their model's

performance actually declined (five ratings increased). Though iterative model refinement is a cornerstone of IML pipelines [13], none of our participants trained a second model due to time constraints or fatigue. However, they shared several ideas for what they would do differently if they were to personalize a new model in the future.

Some participants sought more practical sound classes, either adding *"more [...] things that are important to me"* (P7) or removing less valuable sounds, such as *"the door opening and closing, I don't think I need that"* (P2). Non-expert users commonly believe they will see performance gains by adding more training data indiscriminately [67]; a few participants voiced similar ideas. But others sought to be more selective with their datasets, seeing promise in further data refinement—such as by removing outliers in the clustering chart (P10: *"the little points there, sticking out"*)—or recording additional sound variations to introduce edge cases. For example, P1 was *"curious"* how her model would respond to *"different kinds of doors or different ringtones"*. Participants also sought to adjust their chosen sound classification schemes. Following misclassifications during testing, P12 aimed to resolve *"overlapping sounds"* in his clustering chart, realizing that: *"Shower and the faucet [...] maybe I could combine [those] and have 'water running"'* (P12). Similarly, P4 sought to improve the performance of his "Fridge" class; recalling its two separated clusters, *"[I would] focus on just the 'ice' [dispenser], just because the 'water' [dispenser] is similar enough to the 'Faucet'. [...] [I'm] confusing the model by having two different sounds come out of the same object."*

6 Discussion

Our work advances understanding of how to support DHH users in training personalized sound recognizers by: (1) investigating non-auditory data representations across an end-to-end training cycle for data collection, training data selection, and practical testing; (2) demonstrating how interactive data clustering can support DHH users to reason about audio data, identify outliers, and refine training datasets; and (3) provide insights into DHH users' experiences and perspectives on personalizing a sound recognition pipeline. Our work also reaffirms prior work showing DHH users' preference for waveforms when recording [20]—while expanding on their value when selecting training data and testing—and how training strategies can change through use [50]. Below, we situate our findings in the literature, offer design considerations, and discuss limitations and opportunities for future work.

6.1 Design Considerations for Interactive Sound Recognition Tools

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

Interactive clustering visualizations. We found clustering effectively provides non-auditory feedback on interclass relationships, supporting DHH users' understanding of an audio dataset (a key challenge identified in prior work [20]) as well as their iterative refinement of a training subset. In this way, the visualization enables more active participation in model training —another challenge for DHH users [50]. To better highlight the impact of training data inclusion or exclusion, participants requested visual cues or side-by-side comparisons. Designers should also be mindful of potential overfitting when users rely solely on clustering for training data selection. Future work could investigate how inclusion and exclusion decisions may impact model performance and provide user feedback accordingly. To further mitigate overfitting, encourage users to focus on outliers and overlap and emphasize clustering as a representation of the model's perspective rather than the ground truth of decision boundaries.

Waveforms. Our findings suggest that waveforms are essential for DHH users throughout the IML workflow, and their single dimension of amplitude (loudness) *vs.* time is intuitive for this population. For DHH users to monitor sound input and their soundscape, waveforms should be displayed prominently before and during recording (extending prior work [20]) and when testing models. When selecting training data, waveforms offer a glanceable representation that provides transparency into individual audio examples and adds context for locations within the clustering visualization.

Spectrograms. While spectral features are the standard input for sound recognition models [22], spectrogram visualizations did not offer a significant benefit for DHH users in our study. In contrast to the waveforms' simple vertical amplitude, spectral information depicted by frequency on the vertical and amplitude as color intensity is less intuitive—even confusing—for DHH users. However, spectrograms may offer limited value for in-depth analysis when selecting training data (particularly to experienced users [61]). Other time-frequency visualizations, such as correlograms or pitchograms [9], may offer more value as visual analysis tools for this population and present an opportunity for future work.

Annotating. Our findings suggest that allowing DHH users to provide notes about a sound's production, location, and context aids their understanding and ability to use an IML workflow. Some annotations drove users' exploration of nuanced subcategories within a sound class; designers can proactively support this by suggesting potential subcategories from the start (*e.g.*, generating subcategories for a given sound class via a language model). Highlighting annotated subcategories visually in the clustering (*e.g.*, P11: two discrete clusters for different phone ringtones) can further expose distinctions in the model's decision space, aiding comprehension. Concept decomposition options [52] can streamline clustering insights—either to separate disparate clusters into their subcategories (*e.g.*, P4: fridge → water, ice dispenser) or to combine overlapping classes (*e.g.*, P12: faucet, shower → water running).

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

System format. Participants wanted to capture data on mobile devices but requested flexibility in the device for IML; cloud-based applications can allow users to take a multi-device approach. When adapting IML workflows to mobile formats—the device form factor that is ultimately preferred for daily sound awareness [15]—prioritize waveforms for data collection, clustering throughout training, and waveforms + predictions during testing.

6.2 Future Opportunities for Efficiency and Model Optimization

We found clear limits to the time and effort that DHH users are willing to invest in personalizing sound recognition models, creating a tension with streamlining personalization tasks for efficiency without reducing users' engagement in an interactive training process [1, 13]. Prior work [28] found that Android sound recognition users hoped to spend less than 25 minutes on personalization; training a model with SPECTRA required considerably more time—the allotted hour for most users—and those with time remaining declined to retrain their model. While users found the clustering visualization highly engaging, data selection was a key area to streamline: improved data processing (*e.g.*, automatic segmentation, silence removal) can reduce data cleaning efforts, while automatic outlier detection (*e.g.*, [54]) can highlight atypical examples.

Optimized ML architectures or extensions to pre-trained models can further reduce the effort required for interactive personalization. While SPECTRA's Speech Commands API [62] was well-suited for prototyping interactive training workflows for our study, Jain *et al.*'s ProtoSound is a more optimal architecture for the daily needs of DHH users (*e.g.*, contextual flexibility, open set classification) [28]. Protosound combines few-shot learning with prototypical networks to train custom sound models that outperform comparable architectures (including with DHH users' recordings); however, ProtoSound's "black box" interface lacks audio visualizations and control of the training dataset. As SPECTRA focuses specifically on frontend support, combining it with ProtoSound is a clear next step to improve the baseline performance of users' models, reducing time for model refinement. For example, SPECTRA's clustering visualization integrated within the ProtoSound pipeline could include embeddings generated from the model's internal feature representations, yielding true insight into how the model differentiates sounds and its decision-making process. Finally, our findings also highlight that rather than create a new model, some users feel that adding custom classes to existing sound models is a simpler task. Similar customization features are supported in iOS (tuning existing classes) [2] and Android (adding new classes) [21], but these, too, lack accessible data representations and training insight; supporting the interactive extension of pre-trained models is an opportunity for future work.

6.3 Limitations

Our work has three primary limitations. First, we did not conduct formal analyses of the participants' models and thus cannot definitively quantify the impact of SPECTRA's accessibility features and participants' decisions on model performance. Further, the quality assessment stage was limited to reproducing their model's trained sounds within the system and did not include metrics about the model; as a result, participants' high opinion of their models may have been inflated due to the lack of long-term practical use. Second, while our evaluation presented SPECTRA within a familiar, domestic soundscape, we acknowledge that new challenges with user-driven personalization may emerge in complex acoustic environments with many similar and/or overlapping sounds. Future longitudinal studies are needed to investigate interactive training and deployment within busy or unfamiliar locations—settings where DHH users may most need sound awareness support [15, 20]. Next, while we did not aim to recruit people with ML expertise, five of the 12 participants had hands-on ML experience. Though often not extensive and not in the audio domain, this experience suggests that our participants may not represent the general public, limiting the generalizability of our findings. Finally, we had limited control of the testing environment: our remote evaluation using videoconferencing software led to greater setup and troubleshooting time while reducing opportunities for retraining, experimentation, and/or discussion for some participants. The abbreviated experience may have impacted participants' opinions about their models, and future longitudinal studies can better explore how users' perceptions change through continued use.

7 Conclusion

In this paper, we presented SPECTRA, an interactive system to meet the needs of DHH users when training personalized sound recognition models. We evaluated the prototype in a hands-on model-training session with 12 DHH participants; our findings highlight the potential for interactive clustering and audio visualizations to support accessible exploration and interpretation of an audio dataset, and rich text annotations to prompt varied and realistic data collection. In addition, we explore the unique opportunities and challenges that interactive training poses to DHH users, including the impacts of this experience on users' confidence and understanding of sound recognition systems. Our work provides insights for the design of future tools in this area.

Acknowledgments

We thank Dhruv Jain for his support in resolving early technical hurdles. 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-2140004, and by University of Washington CREATE.

Disclaimer: Generative AI software was used in the preparation of this document for minor improvements to clarity and grammar, and to generate the name/acronym for our system (SPECTRA).

Conflict of interest disclosure: Leah Findlater is also employed by and has a conflict of interest with Apple Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc.

References

[1] 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

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

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

[4] 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

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

[6] 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

[7] 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

[8] 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

[9] 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

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

[11] 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

[12] 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

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

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

[15] Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and Hard-of-hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300276

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

[17] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: Composing Interactive Machine Learning Workflows and Interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 39–53. https://doi.org/10.1145/3472749.3474734

[18] 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

[19] 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

[20] Steven M. Goodman, Ping Liu, Dhruv Jain, Emma J. McDonnell, Jon E. Froehlich, and Leah Findlater. 2021. Toward User-Driven Sound Recognizer Personalization with People Who Are d/Deaf or Hard of Hearing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (June 2021), 1–23. https://doi.org/10.1145/3463501

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

[22] 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.

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

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

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

[26] 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

[27] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People Who Are Deaf and Hard of Hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '18)*. ACM, New York, NY, USA, 81–92. https://doi.org/10.1145/3234695.3236362

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

[29] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, Leah Findlater, and Jon Froehlich. 2019. Exploring Sound Awareness in the Home for People who are Deaf or Hard of Hearing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300324

[30] 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

[31] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. SoundWatch: Exploring Smartwatch-based Deep Learning Approaches to Support Sound Awareness for Deaf and Hard of Hearing Users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2020)*. ACM.

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

[33] 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

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

[35] 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

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

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

[38] 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

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

[40] 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

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

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

[43] 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

[44] Tara Matthews, Janette Fong, F. Wai-Ling Ho-Ching, and Jennifer Mankoff. 2006. Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology* 25, 4 (July 2006), 333–351. https://doi.org/10.1080/01449290600636488

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

[46] 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

[47] 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

[48] 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.

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

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

[51] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 293:1–293:10. https://doi.org/10.1145/3173574.3173867

[52] 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

[53] 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

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

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

[56] 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

[57] Liu Sicong, Zhou Zimu, Du Junzhao, Shangguan Longfei, Jun Han, and Xin Wang. 2017. UbiEar: Bringing Location-independent Sound Awareness to the Hard-of-hearing People with Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (June 2017), 17:1–17:21. https://doi.org/10.1145/3090082 Place: New York, NY, USA Publisher: ACM.

[58] 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

[59] 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

[60] I. R. Summers, M. A. Peake, and M. C. Martin. 1981. Field Trials of a Tactile Acoustic Monitor for the Profoundly Deaf. *British Journal of Audiology* 15, 3 (Jan. 1981), 195–199. https://doi.org/10.3109/03005368109081437

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

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

[63] 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

[64] 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

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

[66] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. 2021. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445306

[67] 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

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

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