A Human-AI Collaborative Approach for Designing Sound Awareness Systems

Jeremy Zhengqi Huang
University of Michigan
zjhuang@umich.edu

Reyna “Wren” Wood
University of Michigan
reynaw@umich.edu

Hriday Chhabria
University of Michigan
hridayc@umich.edu

Dhruv Jain
University of Michigan
profdj@umich.edu

ABSTRACT

Current sound recognition systems for deaf and hard of hearing (DHH) people identify sound sources or discrete events. However, these systems do not distinguish similar sounding events (*e.g.,* a patient monitor beep vs. a microwave beep). In this paper, we introduce HACS, a novel futuristic approach to designing human-AI sound awareness systems. HACS assigns AI models to identify sounds based on their characteristics (*e.g.,* a beep) and prompts DHH users to use this information and their contextual knowledge (*e.g.,* “I am in a kitchen”) to recognize sound events (*e.g.,* a microwave). As a first step for implementing HACS, we articulated a sound taxonomy that classifies sounds based on sound characteristics using insights from a multi-phased research process with people of mixed hearing abilities. We then performed a qualitative (with 9 DHH people) and a quantitative (with a sound recognition model) evaluation. Findings demonstrate the initial promise of HACS for designing accurate and reliable human-AI systems.

CCS CONCEPTS

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

ACM Reference Format:
Jeremy Zhengqi Huang, Reyna “Wren” Wood, Hriday Chhabria, and Dhruv Jain. 2024. A Human-AI Collaborative Approach for Designing Sound Awareness Systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3613904.3642062

1 INTRODUCTION

Past studies indicated that deaf and hard of hearing (DHH) people seek enhanced sound awareness [1, 3]. Current sound awareness systems like SoundWatch [16] identify sound sources and events but often misidentify sounds due to the lack of contextual information and difficulty distinguishing similar sounds (*e.g.,* reporting a smoke alarm while the heart rate monitor beeps in the hospital room) [13]. To address this, we propose a fundamentally novel approach to designing human-AI sound awareness systems, HACS (Human-AI Collaborative Sound Awareness). The underlying idea behind HACS is that AI models recognize and inform DHH users of the sounds based on “how they sound like” (instead of the sound events or sources), and then DHH users leverage their knowledge about the situated contexts to recognize actual sound events. For example, if a DHH user has just finished cooking in the kitchen, and the system reports “liquid flowing,” the user may recognize the sound as a dishwasher running instead of a washing machine, even though the two sounds are similar. Beyond the potential to increase accuracy, this approach drastically increases the agency of DHH users in the sound processing process, enabling them to take over the role of “predictor” to accurately determine sound events.

As a first step of implementing HACS, we articulated a characteristics-based sound taxonomy that categorizes sounds based on “how they sound like.” We began by compiling a list of eighteen source-ambiguous sounds from Audio Set [10] that covered DHH people’s desired sounds [8]. To arrive at Deaf-friendly representations of these sounds, we conducted semi-structured interviews with eight interpreters to understand how these sounds are represented in American Sign Language (ASL). We also invited these interpreters to complete a card sorting task by clustering the 18 sound items into categories based on similarities in their signs. Based on the insights from the semi-structured interviews, card sorting activities, and in-depth discussions with a research team of mixed hearing abilities (2 ASL interpreters, 2 CART writers, 1 DHH researcher, 2 hearing researchers), we articulated an 18-class, characteristics-based, ASL-friendly sound taxonomy.

We conducted two preliminary evaluations of HACS. The first study simulated a scenario where a HACS-based system reported the characteristics of a sound (*e.g.,* beep) that was taking place in a given context (*e.g.,* a kitchen) and prompted DHH participants to recognize sound events based on this information. Participants’ responses provided initial evidence that HACS can help DHH users distinguish similar sound events (*e.g.,* blender *vs.* vacuum cleaner). Moreover, HACS showed the potential to help DHH users identify different “states” of the sounds generated by the same sound source (*e.g.,* door knock *vs.* door slam). For the second evaluation study, we trained a sound recognition model based on the characteristics-based taxonomy to assess whether the sound classes enclosed in the taxonomy could be accurately recognized through algorithmic approaches. We found a near-perfect classification accuracy

CHI ’24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642062

(98.6%) when evaluated on a small dataset, further demonstrating the promise of our approach.

In summary, our work contributes (1) a novel human-AI approach to designing sound awareness systems, (2) an 18-class taxonomy that classifies sounds based on sound characteristics, and (3) two preliminary evaluations demonstrating the initial promise of our approach in accurately recognizing sound events.

2 RELATED WORK

We present background on and situated our work within DHH culture and American Sign Language, sound classification schemes, and state-of-the-art sound awareness solutions.

2.1 DHH Culture and American Sign Language

Deafness is not just audiologically represented. Researchers have represented hearing loss through three models of disability: medical, social, and cultural [6, 24, 32]. While medical and social models emphasize physiological, social, and environmental barriers, the cultural model of deafness embodies a linguistic and cultural group (*i.e.,* DHH Culture). DHH culture is a diverse cultural milieu characterized by an established set of values, norms, behaviors, and languages like American Sign Language (ASL) [6, 25]. ASL is a natural language with linguistic components like syntax and grammar and is capable of expressing complex and abstract ideas, emotions, and narratives in a visual-spatial modality [19, 28, 31]. An important ASL concept relevant to our study is *classifiers*, a morphological system that can represent events and states [12]. Classifiers can represent an entity, describe the size and shape of the objects, and indicate the interactions between objects [12]. These attributes can help describe sound events, as sounds can be interpreted as the interactions of materials in an environment [9]. Our study extends the impacts of ASL to the development of assistive technologies by exploring how everyday sounds can be effectively represented based on sign language interpretations.

2.2 Sound Classification Schemes

Researchers have explored systematic sound classification schemes for decades. We review the four common ones: source-based, interaction-based, signal-based, and hybrid approaches.

Source-based. Early researchers like Schafer pioneered soundscape research and categorized environmental sounds based on the presence of human activities [29]. Following this work, many studies proposed sound classifications based on sound sources across different domains, including urban areas [5, 27], restaurants [20], and geographical locations [14].

Interaction-based. Gaver [9] proposed an "ecological" and interaction-based approach for sound classification based on the material of the sound sources and their physical interactions with the environment. For example, the sound of a waterfall could be described as a large amount of liquid pouting into a pond from high elevation + high-force splash.

Signal-based. The signal-based classification scheme concerns the acoustic signals or audio features [22, 23]. For example, Mitrović et al. [22] classified sounds based on the perceptual properties of sounds, including amplitude and pitch.

Hybrid approach. Many recent sound classification approaches are based on both semantic and signal-based properties of the sound [2, 10]. For example, Audio Set, a 632-class sound taxonomy, categorizes sounds based on both high-level, semantic relations of sound sources (*e.g.,* animals – pets – dogs) and more general sound characteristics like "whir" [10].

Regardless of the approaches, most of the sound classification schemes are based on the auditory perception and cognition of hearing people. An exception includes Rosen's work that probed the representations of sounds in the American Deaf Culture [26]. However, to our knowledge, no research focused on developing sound classifications from a DHH-centric perspective. While developing DHH-centric sound classification schemes seems counterintuitive, many prior studies demonstrated the benefits of sound awareness for DHH people (*e.g.,* helping perform everyday tasks) [3, 8, 15, 16]. To address this gap, we articulated a sound taxonomy that classifies sounds based on sound characteristics depicted in ASL (*e.g.,* ASL signs for whirring and liquid flowing).

2.3 State-of-the-Art Sound Awareness Solutions

The current state-of-the-art sound awareness systems [15, 16, 18] apply a discrete, source-based approach (*i.e.,* identifying discrete sound events like door knock) to classify sounds. However, field studies of these technologies with DHH users [13, 15] showed that the source-based approach is not accurate or reliable enough for everyday use due to several limitations. First, source-based systems may fail to distinguish sound events with similar physical properties (*e.g.,* door knock and footsteps). Second, source-based systems may fail to recognize different "states" of a sound source (*e.g.,* different cycles of washing machines). Third, these systems often lack contextual knowledge, leading to sound feedback that might be context-inappropriate (*e.g.,* recognizing the patient monitor beep as a smoke alarm). AdaptiveSound [7] responded with a feedback-loop system that enabled DHH users to provide feedback on the model output to make it more contextually appropriate, but the challenges in classifying similar sounding sounds remain. This work addresses the above limitations by proposing a novel approach for designing sound awareness solutions that leverages the strengths of both AI models (*i.e.,* pattern recognition) and DHH users (*i.e.,* contextual awareness) while recognizing sound events.

3 THE HUMAN-AI COLLABORATIVE SOUND AWARENESS (HACS) APPROACH

As we described in Section 2.3, prior field studies [13, 15] found that the sound recognition systems that identify discrete sound events or sources were prone to errors due to the AI system's lack of contextual knowledge and the ability to distinguish similar-sounding sounds (*e.g.,* patient monitor *vs.* smoke alarms). This finding motivated us to design a new approach that addresses the above limitations.

We propose HACS, a novel approach for designing human-AI sound awareness systems (Figure 1). The main idea of HACS is to leverage both AI's pattern recognition and DHH users' contextual awareness abilities to achieve sound awareness. The human-AI sound recognition systems based on the HACS approach work as follows:

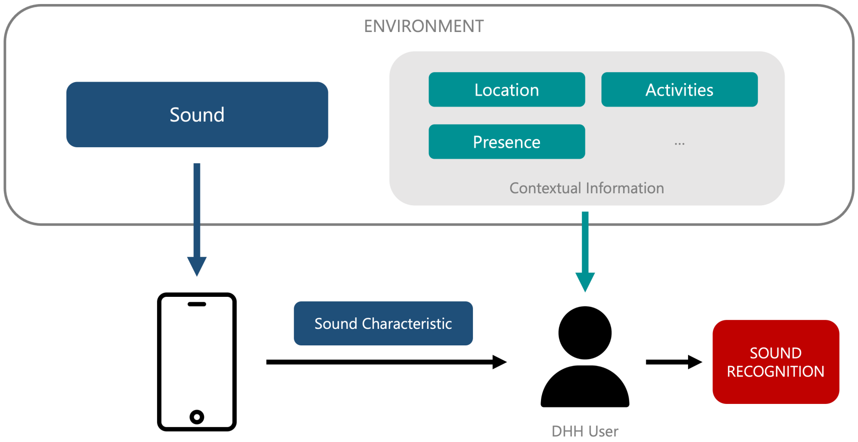

Figure 1: The HACS approach for human-AI collaborative sound awareness.

1. The sound recognition model receives audio signals from the environment, processes the audio, and recognizes its characteristics.
2. The model informs DHH users of the sound characteristics.
3. The DHH users will then use the information enclosed in the sound characteristics and their knowledge about the situated contexts, including the location, activities, etc. (*e.g.,* “I am currently standing in a kitchen cooking”), to recognize the actual sound events.

HACS can address two key shortcomings of the source-based approaches (*i.e.,* identifying discrete sound events or sources). First, by prompting AI models to classify sounds based on characteristics, HACS can bypass AI’s limitations in distinguishing sounds with similar physical properties. Second, HACS incorporates contextual knowledge of DHH users, which can help elicit context-appropriate sound recognition. For example, when DHH users are informed of a “beeping” (sound characteristic) sound in a hospital room (situated context), they are more likely to recognize the sound as a patient monitor beep rather than a microwave beep (sound event). Similarly, suppose a DHH user just finished cooking and receives a “liquid flowing” sound notification. In that case, they may recognize the sound event as a dishwasher running rather than a washing machine, even though the two sound events are similar.

4 CONSTRUCTING A CHARACTERISTICS-BASED SOUND TAXONOMY FOR HACS

We articulated a characteristics-based taxonomy as the first step for implementing HACS. The design goals for this taxonomy are twofold. First, the sound classes presented in the taxonomy should support automatic classification (step 1 of the HACS). Second, the information enclosed in the taxonomy should be informative and intuitive enough to help DHH people make sense of the sound events (steps 2 and 3 of the HACS).

Table 1: Demographics of ASL interpreters for the formative study

Participant ID	Gender	Age	Years of Experience
I1	Female	26	5 years
I2	Female	50	13 years
I3	Male	21	1 year
I4	Male	30	8 years
I5	Female	23	1.5 years
I6	Female	26	5.5 years
I7	Male	34	11 years
I8	Female	47	29 years

4.1 Formative Study Methods

Participants: As an integral part of DHH culture, ASL can bridge the gap between DHH people and sounds. This property makes ASL signs a fitting medium for describing sound characteristics in a way that is closely aligned with DHH culture. Therefore, we recruited 8 ASL interpreters through online study ads, social media (*e.g.,* Reddit posts), emails, and snowball sampling (see Table 1). The average age of these participants was 32.1 years old (*SD*=10.9, *range*=21-50). The average years of experience was 9.25 (*SD*=9.0, *range*=1-29). All interpreters were U.S. residents and had experience working with DHH people professionally.

Procedure: To begin constructing our characteristics-based taxonomy, the first and second authors, who are hearing, independently selected source-ambiguous sounds from the Audio Set ontology [10] that could comprehensively represent sounds in real life. For example, the first author considered the “whir” sound from the Audio Set ontology to represent appliances running (*e.g.,* washers and dryers) and the “thump” sound to represent dull objects like books dropping to the floor. The authors then met to construct a singular list of sounds by sorting, splitting (*e.g.,* separating “thump” and “tap” into independent categories), and combining (*e.g.,* “rip” and “tearing” into a single category) the sound categories. This process was guided by DHH people's desired sounds elicited by Bragg et al. and Findlater et al.'s large-scale surveys [3, 8]. More specifically, we ensured that the characteristics of all desired sound sources by DHH people in past work are covered by our list of sounds (*e.g.,* we included "whir" of the rotatory motor and "beep" sounds to cover both classes of sounds emitted by "microwave," a commonly desired sound source by DHH people in prior work). This resulted in a list of 18 source-ambiguous sound items. We then acquired audio files for these 18 classes by searching their labels (*e.g.,* "whir") on FreeSound [34].

To understand how our list of 18 sound items can be represented in ASL and to further refine our list, we conducted semi-structured interviews with eight ASL interpreters in Spoken English via Zoom. We first asked participants to complete a brief background form to collect their demographic information and experience with sign language. We then asked 15 questions about (1) interpreters' experience and contexts of working with DHH clients and (2) sign language interpretations of everyday sounds. Finally, we invited these interpreters to engage in a card-sorting task using a FigJam board with the 18 sound items, each labeled S1 to S18; see Appendix A3 for the list. For each sound item, the interpreters listened to the clips once, demonstrated or described the signs, and were instructed to freely move and cluster the sound items into categories based on how similar these sounds could be. We played only 5 seconds of each sound clip to ensure that our study stayed within the time limit. Interpreters could replay the clip if needed. We allocated 30 minutes for this task. Seven interpreters completed the task within this time, with one taking five extra minutes. The first and second authors observed and recorded the task as edited FigJam files. All interpreters shared their screens while completing the task.

Analysis: Our formative study data consisted of the transcripts of eight interview sessions obtained from real-time captioners and eight edited FigJam files. For the transcripts, we used Braun and Clarke's six-phase approach [4]. The first author skimmed and familiarized with the data (step 1) and discussed with the research team to generate an initial codebook (step 2). The first author then walked through the data in detail and iteratively applied the codes to the data while refining the codebook. The final codebook had a 3-level hierarchy: 6 first-level, 17 second-level, and 63 third-level codes. The second author independently applied the codes based on the final codebook (step 4). We calculated the interrater reliability between two coders using the ReCal 2 package [33] and resolved the disagreements among coders. The average Krippendorff's alpha value was 0.696, and the raw agreement was 84.3%. Finally, we organized the first-level themes (step 5) and constructed our narratives accordingly (step 6). We have attached our final codebook as supplementary material.

We also performed a cluster analysis on the participants' responses in the card sorting task, which consisted of eight edited FigJam files. We first walked through the individual files and logged the clusters formed by participants. For example, if a participant formed a cluster that contained S7, S10, S12, we noted "<S7, S10, S12>". After all clusters were logged, we listed all two-item pairs within individual clusters. Using the above example, the three two-item pairs would be <S7, S10>, <S7, S12>, and <S10, S12>. We then constructed an 18x18 similarity matrix to visualize the co-occurrences (see Appendix A1). We considered sound pairs with 50% or more agreement (four or more interpreters) "signed similarly." The authors then further discussed the results with two ASL interpreters within our research team, during which we covered: (1) what sound items should be merged and (2) what other sounds should be included. Upon consensus on the sound classes, we invited the two interpreters to generate the "glosses" (approximate written descriptions of the ASL signs, see Table 2) for each sound class. We also invited two CART writers to generate the English captions for the sound classes after listening to the corresponding sound clips (*e.g.,* [LIQUID FLOWING]). These captions were later used as the labels for the taxonomy's sound classes (see Table 2).

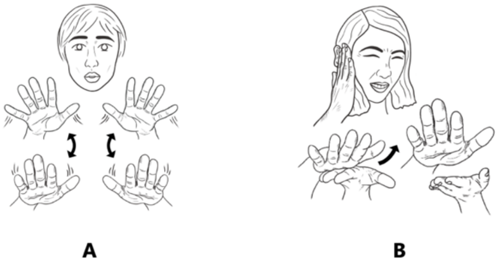

Figure 2: The signs demonstrated by interpreters who used classifiers + NMM combinations to represent sounds.

4.2 Findings from the Formative Study

Findings: During the interview, interpreters reported a diverse set of techniques for describing sound characteristics in ASL without disclosing sources, including classifier construction and/or Non-Manual Markers (NMM; N=6) and listing possible sounds (N=3). For example, I7 demonstrated that "rumble" sounds like "jet flying overhead" could be signed with classifiers and NMMs like "widening eyes," the subtle "blowing air" expression, and the "up-and-down" hand movements (Figure 2A). Similarly, "screechy or squeaky" sounds could be described with both hands, with the CL-B classifier indicating a motion of "bottom of an object rubbing against a surface" and the NMM of "harsh sounds" (I1; Figure 2B). These patterns were reflected during the card sorting task, as for most sounds, the signs demonstrated by the interpreters were similar for the same sound with minor differences in classifiers (*e.g.,* different handshapes for representing objects). For example, all interpreters demonstrated the "thump" sounds as objects falling onto the ground. However, the handshapes used to represent the objects were different (*e.g.,* CL-S *vs.* CL-5). Three interpreters also stated that indicating possible sounds, or "*reference points*" (I3), could help DHH people make sense of the sound events based on the contexts.

During the card sorting task, all interpreters produced similar signs for the same sound with minor differences in classifiers (*e.g.,* CL-S *vs.* CL-5 when signing the "thump" sound). The cluster analysis on the card sorting task responses showed that four sound pairs had similar signs: *tap* – *knock*, *beep* – *tap*, *breaking* – *splash*, and *whir* – *rolling*. Per interpreters' recommendations, we merged the *tap* and *knock* into one sound class due to similar signs and did the same for *whir* and *rolling*, leading to a list of 16 sound items with

Table 2: The Sound Taxonomy designed for the Human-AI Collaborative Sound Awareness Framework

Class	Class Label	ASL Sign Description	Examples
C1	Liquid Flowing	**WATER** running down from top	Water coming out of faucet, river flowing, sizzling oil
C2	Machine Humming	**MACHINE** running (NMM: puffed cheek)	Car engine, dryer
C3	Shatter	Object **(CL-S)** fall and break suddenly into pieces	Glass bottle breaks, ceramic breaks apart
C4	Fracture	**STICK (CL-G for both hands)** breaks apart	Breaking branches, chopping wood
C5	Rip/Tear	**PAPER** being teared apart	Ripping clothes, peeling tape
C6	Splash	Object **(CL-S)** strike or fall into liquids	Jumping into water, Stepping into mud
C7	Screech	Indicating a harsh sound from the contact of two surfaces **(CL-B)**	Sudden brake (cars, bicycles), nail on a chalkboard
C8	Blender	**BLENDER**	Juicer, blender, coffee grinder
C9	Electrical Buzzer	**BUZZER**	Neon lights, basketball court buzzer
C10	Beep	**BEEP**	Car horn, microwave beep, fire alarm
C11	Knock/Tap	Knocking **(CL-S)** on a surface	Door knock, raindrops hitting on the window
C12	Thump	Dull object **(CL-5)** falls on and hit the surface **(CL-5)**	Footsteps, book falling onto the ground
C13	Bam/Bang	Sound and vibration of a hard blow	Gunshot, fireworks, thunder
C14	Scrape	Object **(CL-C)** Scrape or scratch on a surface **(CL-B)**	Chair being dragged,
C15	Ding/Clink	**BELL** / *downward* CL-A hits CL-B and reverberates	Bell, Toast with wine glasses
C16	Squeal/Shriek	**SCREAM**	Scream, pig squeals, birds
C17	Whoosh	Object **(CL-3)** passing with high speed (NMM: thick cheeks blow air out)	Car passing, strong wind
C18	Crumple	Crush things into wrinkles (NMM: grind teeth)	Aluminum foil, candy wrappers

distinct signs. Further discussions elicited two source-ambiguous sounds from the Audio Set's *source-ambiguous sound* category [10] that are common in real life and have distinct signs: *ding* (*e.g.,* bike bells and wine glasses) and *whoosh* (*e.g.,* wind, a car passing with high speed) – resulting in a list of 18-class, ASL-friendly sound taxonomy that categorizes sounds based on how ASL describes their characteristics.

4.3 The Characteristics-Based Sound Taxonomy

The novel sound taxonomy classifies sounds based on the ASL descriptions of their characteristics (*e.g.,* "machine humming"). The taxonomy, outlined in Table 2 below, contains four fields: Class Code, Class Labels, ASL Sign Descriptions, and Examples. As we mentioned in Section 4.1, the class labels were derived from the captions generated by CART writers. The *ASL Sign Description* field delineates how the sound classes are described with ASL signs and contains classifiers (*e.g.,* **CL-G**), generic objects with dedicated signs (*e.g.,* **PAPER**), and/or non-manual markers (NMM). The *Examples* field serves as "reference points" and helps DHH people understand the kind of sound events each sound class represents (*e.g.,* the *Crumple* sound class can be produced by candy wrappers or aluminum foils).

The information enclosed in the sound classes should be interpreted as: the [Class Label] sound can be represented in ASL by following [ASL Description], and the sound events like [Examples] belong to this class. For example, the "Thump" sound can be represented in ASL through the signs that describe dull objects (**CL-5**) falling on and hitting the surface (**CL-5**), and sound events like footsteps and books falling onto the ground fall into this category.

In summary, we propose HACS, a novel approach for designing sound awareness systems. In HACS, AI models will be trained to identify the sound characteristics instead of discrete sound events or sources, and DHH users can leverage this information and their knowledge about the situated contexts to recognize the sound events. As a first step for implementing HACS, we articulated an 18-class sound taxonomy that classifies sounds based on the sound characteristics depicted in ASL.

5 PRELIMINARY EVALUATIONS OF HACS

We conducted two preliminary evaluation studies to assess the feasibility of the HACS approach. Our objectives were twofold. First, we examined if the information enclosed in HACS' sound classes could help DHH individuals identify sound events relevant to the specific contexts using their contextual understanding. Second, we evaluated whether our taxonomy's classes could be classified accurately using algorithmic approaches. We describe these evaluations in detail below.

Table 3: PE1 participants' background information. SimCom stands for "simultaneous communication," a communication method where people use spoken language at the same time. PMOC stands for "preferred mode of communication."

PID	Gender	Age	Identity	Hearing loss	ASL Exp.	PMOC
P1	Male	21	Deaf	Profound	2 years	Sign Language (D) Sign Language (H)
P2	Female	20	deaf	Moderate	2 years	Sign Language (D) SimCom (H)
P3	Female	33	deaf	Profound	32 years	Sign Language (D) Writing (H)
P4	Female	30	Deaf	Moderate	2 years	Sign Language (D) Sign Language (H)
P5	Female	47	Deaf	Profound	45 years	Sign Language (D) Writing/Texting (H)
P6	Male	72	Deaf	Severe	45 years	Sign Language (D) Verbal (H)
P7	Female	28	Deaf	Profound	26.5 years	Sign Language (D) Writing (H)
P8	Female	59	Deaf	Profound	30+ years	Sign Language (D) Verbal (H)
P9	Female	37	Deaf	Profound	Whole life	Sign Language (D) Verbal (H)

5.1 Preliminary Evaluation 1: Online Simulation of HACS-based Systems

The goal of PE1 was to evaluate whether the information enclosed in the characteristics-based taxonomy could help DHH individuals identify sound events using the contextual knowledge. This evaluation tested steps 2 and 3 of the HACS workflows we described in Section 3.

Participants: PE1 sessions were conducted by the first, second, and fourth authors, one of whom is DHH. We also recruited an ASL interpreter and a real-time captioner to facilitate communication for all sessions. We proceeded with sessions once we received the participants' consent with IRB-approved consent forms. At the beginning of the session, we asked the participant to complete a background form asking about their demographic information (see Table 3). The average age of these participants was 38.6 years old (*SD*=17.6, *range*=20-72). The average years of experience signing ASL was 24.6 years (*SD*=18.0, *range*=2-45).

Study setup and procedure: Before the PE1 sessions, we acquired 18 sound clips from FreeSound [34] by searching with the sound class labels (*e.g.,* "shatter" and "thump"). We also prepared a list of contexts from three categories (*i.e.,* indoor–home, indoor–public spaces, and outdoor; Figure 3), which would be presented to participants. Participants were briefed about the taxonomy to understand the sound classes better. During the sessions, we created an online simulation via Zoom, where a HACS-based sound awareness system identified a sound class from the characteristics-based taxonomy and informed DHH users, who possessed knowledge of their situated contexts. Specifically, for each sound class, we followed the below steps (see Figure 4):

1. We played the sound clip while ASL interpreters signed the sound based on the "ASL Sign Description" field of the taxonomy.
2. We presented the DHH participants with the corresponding sound information, including "Class Label," "ASL Sign Description," and "Examples." We also selected one context from each of the three categories (a total of three) and presented these contexts to the DHH participant. The context selection within the categories rotated for participants. For example, we presented P4 with "kitchen" and P5 with "living room" for the home category.
3. Participants were asked to imagine being in the given context and to suggest possible sound events that were occurring.

After the sessions, we asked DHH participants for overall feedback about HACS and the characteristics-based taxonomy.

Data Analysis: Our PE1 data consisted of the transcripts of nine study sessions obtained from the real-time captioner, and a list of sound events participants inferred based on sound and contextual information. For the transcripts, we used the same analysis approach as that for the Formative Study (Braun and Clarke's six-phase approach [4]; Section 4) to analyze interview transcripts, resulting in a Krippendorf's alpha of 0.692 and raw agreement of 84.8% between the two coders. We attached the final codebook as supplementary materials and the list of participants' inferred sound events in the Appendix A2.

Findings: Based on the responses, we elicited a list of inferred sound events across different contexts for each sound class (see Appendix A2). Here, we highlight several important insights that demonstrate the flexibility and robustness of the HACS approach in supporting sound awareness for DHH people.

First, DHH participants successfully identified different sound events within the same sound class across various contexts. For instance, P9 recognized the "Beep" sound (C10) as a car honk on a "busy street" and as a "ping for the pick-up orders" in a restaurant. Another example is P6 recognizing the "Blender" sound (C6) as a

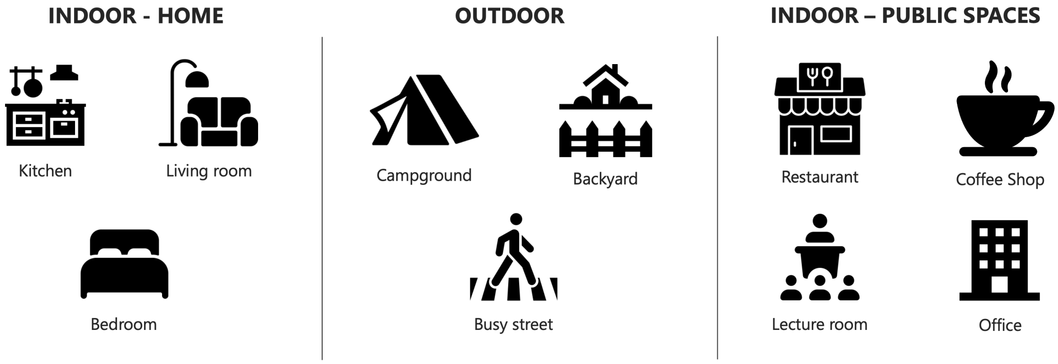

Figure 3: The collection of contexts we used during the Preliminary Evaluation 1.

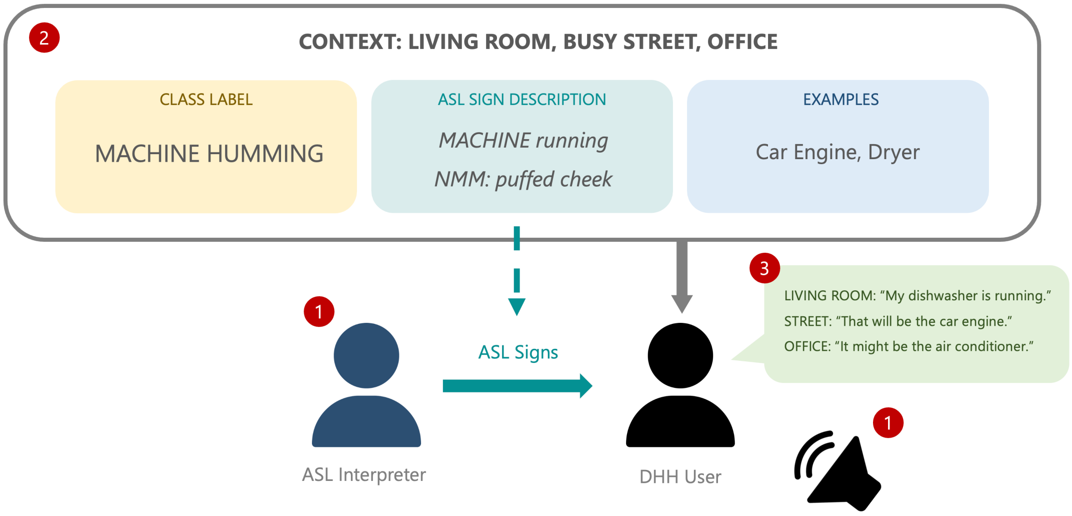
2
CONTEXT: LIVING ROOM, BUSY STREET, OFFICE
CLASS LABEL
MACHINE HUMMING
ASL SIGN DESCRIPTION
MACHINE running
NMM: puffed cheek
EXAMPLES
Car Engine, Dryer
3
LIVING ROOM: "My dishwasher is running."
STREET: "That will be the car engine."
OFFICE: "It might be the air conditioner."
1
ASL Signs
ASL Interpreter
DHH User
1

Figure 4: The setup for Preliminary Evaluation 1: (1) We played the audio clip, and the ASL interpreter signed the sound using the "ASL Sign Description" field; (2) We presented sound class information from the characteristics-based taxonomy and the contexts; (3) We asked the DHH participant to infer sound events based on the information listed in (2).

coffee grinder in a coffee shop and as a vacuum cleaner in a living room. These responses indicated that HACS could help DHH users appropriately interpret sound classes and apply them to recognize sound events in specific contexts.

Second, participants linked one sound source to different sound classes in two ways:

1. Different states or behaviors of the same source. For example, the sounds of "door knock" and "door slam" were identified by P4 as belonging to different classes in a living room context (*i.e.,* Knock/Tap vs. Bam/Bang). In another instance, on a busy street, the "Squeak/Screech" sound was interpreted as the car "screeching to a halt" (P9), and a "Bang/Bam" sound could mean a car accident (P5, P7).
2. Variations in sound sources. For example, P4 and P7 identified the sound of "microwave done" using the "Beep" class, while P9 identified the same event using "Ding/Clink." This reflected participants' experience with different appliance models.

Post-session, we asked participants for feedback about HACS and the taxonomy. The feedback was generally positive, with many finding the ASL information helpful for understanding sound events (N=7). Moreover, P8 noted the adaptability of HACS, stating that this approach could help her learn the differences between the "*windows rolling down*" sound now *vs.* "*50 years ago*". A few participants (N=3) expressed some concerns about the need to expand the taxonomy (P9) and the potential learning curve associated with using it (P6).

5.2 Preliminary Evaluation 2: Sound Classification Experiment

For PE2, we focused on step 1 of the HACS workflow, where sound recognition models process the audio signal and identify the sound characteristics. Specifically, we trained and evaluated a sound classification model based on our characteristics-based taxonomy.

We first compiled our dataset. We downloaded sound clips for the 18 sound classes in our taxonomy from *FreeSound* [34], an online corpus of high-quality, labeled sound effects. All downloaded clips were converted to a single format (16Khz, 16-bit, mono), and silences greater than one second were removed, resulting in 9.8 hours of recordings and a total of 35,280 clips (1,960 per class). We divided our recordings into a train and a test set, with 80% and 20% split, respectively.

To generate input features for our model, we segmented each clip into one-second segments and computed short-time Fourier Transforms using a 25ms sliding window and 10ms step size (frequency range 20Hz to 8000Hz), yielding a 96-length spectrogram. We then converted our linear spectrogram into 64-bin log-scaled Mel spectrogram and generated a 100 x 64 input for every second of audio. To these log-Mel spectrograms, we applied Cepstral Mean and Variance Normalization (CMVN) [30].

To train our model, we adopted a transfer learning approach commonly used for sound classification (*e.g.,* [15, 16, 18]). We downloaded a pre-trained VGG-16 CNN model [11], replaced the last fully connected layer with a fresh layer (using a sigmoid activation function), and finetuned the model on our training set. For training, we used a cross-entropy loss function with an Adam optimizer [17].

We evaluated our model using a clip-level prediction. Specifically, we aggregated the classification confidences for each one-second prediction across the entire clip and returned the top prediction. We found that our model returned a near-perfect accuracy of 98.6% on our test set, demonstrating the potential feasibility of training accurate sound recognition models based on our characteristics-based taxonomy.

6 DISCUSSION, LIMITATIONS, AND FUTURE WORK

We propose HACS, a novel approach for designing human-AI sound awareness systems. HACS uses AI to identify sound characteristics and prompts DHH users to use this information and their knowledge about the situated contexts (*e.g.,* location, activities, human presence, etc.) to identify sound events. HACS addresses previous systems' limitations, especially in distinguishing similar sounds [13].

As a first step for implementing HACS, we articulated a characteristics-based sound taxonomy (Table 2). This taxonomy classifies sounds based on how sound characteristics are depicted by ASL signs, which reflects multiple dimensions of sound characteristics, including the interaction of objects and materials [9], the mechanics of the sound (*e.g.,* continuous *vs.* discrete) [2, 22], and the affective properties (*e.g.,* pleasantness) [2]. Moreover, the inclusion of ASL aligns the characteristics-based taxonomy with DHH culture. For future HACS-based sound awareness systems, the *ASL Sign Descriptions* field can be visualized as animations like Figure 2 (*e.g.,* on a mobile device or a watch).

In two preliminary evaluations, HACS showed potential in helping DHH users recognize contextually appropriate sound events and distinguish sounds with similar physical properties. It also proved adaptable to varying sound environments and was able to help DHH users recognize different sounds produced by the same kind of sound source. PE2 confirmed the feasibility of automatic recognition of the sound classes.

HACS' customizability and adaptability open new possibilities for designing human-AI sound awareness systems in many domains. For example, HACS-based sound recognition systems may allow DHH users to personalize them by assigning labels to sound classes across various contexts. For online videos, HACS can also be applied to overcome the challenges in captioning ambiguous non-speech sounds [1]. It can also be baked into customizable interfaces that visualize the non-speech sounds (*e.g.,* ARAO [21]).

Our study has several limitations, including the reliance on ASL, the need for adaptation to other sign languages (*e.g.,* Indo-Pakistani Sign Language, Chinese Sign Language), and the preliminary nature of evaluations. HACS also does not address the recognition of overlapping (co-occurring) sounds, an open research area. Furthermore, we do not claim that our taxonomy is exhaustive or will work as-is for all users. For example, our taxonomy does not cover music or melodic sound patterns. However, we ensured that all commonly desired sounds by DHH people in past work were covered, and music was not one of them. We welcome future work that further validates and expands our taxonomy. Finally, HACS may only suit some DHH individuals, but its flexibility offers a promising avenue for customizable sound recognition systems.

ACKNOWLEDGMENTS

We thank Yuni Park for helping with the sketches for ASL representations of sounds. We also thank our participants for their valuable insights.

REFERENCES

[1] Oliver Alonzo, Hijung Valentina Shin, and Dingzeyu Li. 2022. Beyond Subtitles: Captioning and Visualizing Non-speech Sounds to Improve Accessibility of User-Generated Videos. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (*ASSETS '22*), October 22, 2022, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–12. . https://doi.org/10.1145/3517428.3544808

[2] Oliver Bones, Trevor J. Cox, and William J. Davies. 2018. Sound Categories: Category Formation and Evidence-Based Taxonomies. *Front. Psychol.* 9, (July 2018), 1277. https://doi.org/10.3389/fpsyg.2018.01277

[3] 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*, October 23, 2016, Reno Nevada USA. ACM, Reno Nevada USA, 3–13. . https://doi.org/10.1145/2982142.2982171

[4] Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide.* SAGE Publications.

[5] A. L. Brown, Jian Kang, and Truls Gjestland. 2011. Towards standardization in soundscape preference assessment. *Applied Acoustics* 72, 6 (May 2011), 387–392. https://doi.org/10.1016/j.apacoust.2011.01.001

[6] Anna Cavender and Richard E. Ladner. 2008. Hearing Impairments. In *Web Accessibility: A Foundation for Research*, Simon Harper and Yeliz Yesilada (eds.). Springer, London, 25–35. https://doi.org/10.1007/978-1-84800-050-6_3

[7] Hang Do, Quan Dang, Jeremy Zhengqi Huang, and Dhruv Jain. 2023. AdaptiveSound: An Interactive Feedback-Loop System to Improve Sound Recognition for Deaf and Hard of Hearing Users. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility*, October 22, 2023, New York NY USA. ACM, New York NY USA, 1–12. . https://doi.org/10.1145/3597638.3608390

[8] 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*, May 02, 2019, Glasgow Scotland Uk. ACM, Glasgow Scotland Uk, 1–13. . https://doi.org/10.1145/3290605.3300276

[9] William W. Gaver. 1993. What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology* 5, 1 (March 1993), 1–29. https://doi.org/10.1207/s15326969eco0501_1

[10] 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)*, March 2017, New Orleans, LA. IEEE, New Orleans, LA, 776–780. . https://doi.org/10.1109/ICASSP.2017.7952261

[11] 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, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017. 131–135. . https://doi.org/10.1109/ICASSP.2017.7952132

[12] Joseph Hill, Diane Lillo-Martin, and Sandra Wood. 2018. *Sign Languages: Structures and Contexts*. Routledge, London. https://doi.org/10.4324/9780429020872

[13] Jeremy Zhengqi Huang, Hriday Chhabria, and Dhruv Jain. 2023. "Not There Yet": Feasibility and Challenges of Mobile Sound Recognition to Support Deaf and Hard-of-Hearing People. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility*, October 22, 2023, New York NY USA. ACM, New York NY USA, 1–14. . https://doi.org/10.1145/3597638.3608431

[14] Lingjiang Huang and Jian Kang. 2015. The sound environment and soundscape preservation in historic city centres—the case study of Lhasa. *Environ Plann B Plann Des* 42, 4 (July 2015), 652–674. https://doi.org/10.1068/b130073p

[15] 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*, April 21, 2020, Honolulu HI USA. ACM, Honolulu HI USA, 1–12. . https://doi.org/10.1145/3313831.3376758

[16] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. SoundWatch: Exploring Smartwatch-based Deep Learning Approaches to Support Sound Awareness for Deaf and Hard of Hearing Users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, October 26, 2020, Virtual Event Greece. ACM, Virtual Event Greece, 1–13. . https://doi.org/10.1145/3373625.3416991

[17] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* (December 2014). Retrieved December 11, 2023 from https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8

[18] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, October 11, 2018, Berlin Germany. ACM, Berlin Germany, 213–224. . https://doi.org/10.1145/3242587.3242609

[19] Scott K. Liddell. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511615054

[20] PerMagnus Lindborg. 2016. A taxonomy of sound sources in restaurants. *Applied Acoustics* 110, (September 2016), 297–310. https://doi.org/10.1016/j.apacoust.2016.03.032

[21] Lloyd May, So Yeon Park, and Jonathan Berger. 2023. Enhancing Non-Speech Information Communicated in Closed Captioning Through Critical Design. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (*ASSETS '23*), October 22, 2023, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–14. . https://doi.org/10.1145/3597638.3608398

[22] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. 2010. Chapter 3 - Features for Content-Based Audio Retrieval. In *Advances in Computers*. Elsevier, 71–150. https://doi.org/10.1016/S0065-2458(10)78003-7

[23] T. Nakatani and HIroshi G. Okuno. 1998. Sound Ontology for Computational Auditory Scence Analysis. July 01, 1998. . Retrieved August 30, 2023 from https://www.semanticscholar.org/paper/Sound-Ontology-for-Computational-Auditory-Scence-Nakatani-Okuno/c81832ddcaba13f595510b8338f40fabf535ebbb

[24] Michael Oliver. 1996. *Understanding Disability*. Macmillan Education UK, London. https://doi.org/10.1007/978-1-349-24269-6

[25] Carol Padden and Tom Humphries. 1990. *Deaf in America: Voices from a Culture*. Harvard University Press, Cambridge, MA.

[26] R. S. Rosen. 2007. Representations of Sound in American Deaf Literature. *Journal of Deaf Studies and Deaf Education* 12, 4 (April 2007), 552–565. https://doi.org/10.1093/deafed/enm010

[27] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM international conference on Multimedia* (*MM '14*), November 03, 2014, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1041–1044. . https://doi.org/10.1145/2647868.2655045

[28] Wendy Sandler and Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139163910

[29] R. Murray Schafer. 1993. *The Soundscape*. Retrieved September 1, 2023 from https://www.simonandschuster.com/books/The-Soundscape/R-Murray-Schafer/9780892814558

[30] O. M. Strand and A. Egeberg. 2004. Cepstral mean and variance normalization in the model domain. 2004. . Retrieved September 14, 2023 from https://www.semanticscholar.org/paper/Cepstral-mean-and-variance-normalization-in-the-Strand-Egeberg/0de27e275803a000babcfa5c06c0683ee1df76e0

[31] Clayton Valli and Ceil Lucas. 2000. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press.

[32] A. M. Young. 1999. Hearing parents' adjustment to a deaf child-the impact of a cultural-linguistic model of deafness. *Journal of Social Work Practice* 13, 2 (November 1999), 157–176. https://doi.org/10.1080/026505399103386

[33] 2010. ReCal2: Reliability for 2 Coders – Deen Freelon, Ph.D. Retrieved September 11, 2023 from http://dfreelon.org/utils/recalfront/recal2/

[34] Freesound - Freesound. Retrieved September 11, 2023 from https://freesound.org/

APPENDIX

A1 - THE SIMILARITY MATRIX FROM THE CLUSTER ANALYSIS

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
S1		0	1	1	0	4	1	1	1	1	3	1	1	2	0	1	1	0
S2	0		1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
S3	1	1		1	0	1	1	0	3	2	1	0	0	0	0	2	1	0
S4	1	0	1		3	1	0	0	0	0	0	0	0	1	1	3	0	1
S5	0	0	0	3		1	0	0	0	0	1	0	0	3	1	1	0	1
S6	4	0	1	1	1		2	0	1	1	5	1	0	2	0	1	1	0
S7	1	0	1	0	0	2		2	1	4	1	1	0	0	0	0	1	1
S8	1	0	0	0	0	0	2		0	2	0	3	1	1	0	0	0	1
S9	1	0	3	0	0	1	1	0		3	1	0	0	0	0	0	1	0
S10	1	0	2	0	0	1	4	2	3		1	1	0	0	0	0	1	1
S11	3	0	1	0	1	5	1	0	1	1		2	0	2	0	0	1	0
S12	1	0	0	0	0	1	1	3	0	1	2		1	1	0	0	0	1
S13	1	0	0	0	0	0	0	1	0	0	0	1		1	1	0	0	1
S14	2	0	1	0	3	2	0	1	0	0	2	1	1		1	0	0	0
S15	0	0	0	1	1	0	0	0	0	0	0	0	1	1		0	0	0
S16	1	1	2	3	1	1	0	0	0	0	0	0	0	0	0		0	2
S17	1	1	1	0	0	1	1	0	1	1	1	0	1	0	0	0		0
S18	0	0	0	1	1	0	1	1	0	1	0	1	1	0	0	2	0	

A2 – A LIST OF SOUND EVENTS IDENTIFIED BY PARTICIPANTS IN PE1

During Preliminary Evaluation 1, we prompted participants to suggest possible sound events given the sound information from the characteristics-based taxonomy and contextual information and noted these sound events. We compile a list of participants' suggested sound events and present it below.

Class #	Class Label	Sound Events
C1	Liquid Flowing/Running	Water coming out of faucet (P6), water running (P4, P7, P8, P9), waterfall (P7), rain (P9), river flowing (P9)
C2	Machine Humming	Car engine running (P5, P6, P7, P8), washer (P4, P5, P8, P9), dryer (P5), lawn mower (P6, P7), fan (P6), cotton mill (P6)
C3	Shatter	Fragile object hits the floor (P5), glass bottle breaking (P6, P7, P8), dish breaks (P6), window breaking (P7)
C4	Fracture	Plastic rod breaks (P5), Wood chopped (P6), Twig snaps (P7), car broken (P9)
C5	Rip/Tear	Paper being ripped (P4, P6, P7, P8, P9), pants ripping (P4), backpack unzipping (P4, P8), opening a package (P7)
C6	Splash	Water spilled (P4, P6, P8), waterfall (P6), person jumping into water (P7, P8), heavy rain (P7), drop things into water (P9)
C7	Squeak/Screech	Whistle (P4), bird calling (P6), feedback from microphone (P7), writing on chalkboard (P9), car screeching to a halt (P9), plane taking off (P9)
C8	Blender	Blender (P4, P5, P6, P7, P8, P9), coffee grinder (P4, P5, P6), garbage disposal (P5), vacuum cleaner (P6)
C9	Electrical Buzzer	Alarm (P4, P5, P7), electric stove (P4), Razor (P5), dryer done (P6), phone buzzing (P9)
C10	Beep	Car honk (P4, P5, P7, P8, P9), microwave done (P4, P7), EKG (P5), phone beep (P5), oven beep (P8), pings for pick-up orders in boba shop (P9)
C11	Knock/Tap	Door knock (P4, P5, P6, P7, P8), knocking on the countertop (P4), footsteps (P5), settling things down on table (P8), object falls on the pavement (P9)
C12	Thump	Ball bounce (P4), footsteps (P4, P6, P7, P8), book falling on the floor (P5, P7), tree falls (P8)
C13	Bam/Bang	Gunshot (P4, P6, P8), door slam (P4), bomb (P5), car accident (P5, P7), fireworks (P7), throwing hammer at wood (P9)
C14	Scrape/Scratch	Chalk on a chalkboard (P5), scratching on table (P5), moving furniture (P6), scratching on a furniture (P4, P5, P7), scrape on a wok (P8), scratching one's face (P9)
C15	Ding/Clink	Timer done (P4), wind chimes (P4), toasting with wine glasses (P4), bell (P5, P6, P7), messaging notification (P8), cash register (P8), microwave done (P9)
C16	Squeal/Shriek	Human scream (P4, P6), animals (P4, P8), baby crying (P5, P8), mouse (P5), bird calls (P6), computer not working (P9)
C17	Whoosh	Car passing (P4, P5, P6, P7, P8), wind blowing (P5, P9), airplane flying overhead (P7), person walking by (P8), motorcycle passing (P9)
C18	Crumple	Plastic bags crumpling (P4), paper crumpling (P5, P8), candy wrapper (P6), walking on piles of leaves (P7), foil (P8), crush food into pieces (P9)

A3 – SOUND LABELS AND CORRESPONDING CODE

Sound Code	Sound Labels
S1	Beep
S2	Crumpling
S3	Sizzle
S4	Grinding
S5	Whir
S6	Knock
S7	Breaking
S8	Bang
S9	Liquid Flowing
S10	Splashing
S11	Tap
S12	Thump
S13	Scraping
S14	Rattle
S15	Rolling
S16	Buzz
S17	Tearing
S18	Squeak