

SoundTrace: Integrating Temporal Context and Episodic Memory for Real-Time Sound Recognition

DHRUV JAIN*, University of Michigan, USA

JASON MILLER, University of Michigan, USA

Environmental sound recognition systems have become increasingly capable, yet they often operate in context-free modes that ignore temporal continuity, environmental patterns, and user feedback. We present SoundTrace, a real-time sound recognition system that integrates temporal context and episodic memory to support adaptive, interpretable inference in everyday environments. SoundTrace augments a neural audio classifier with lightweight memory structures that store symbolic event traces, estimate scene-time and short-range sequence priors, and update those priors through user feedback. During inference, these memory-derived priors are retrieved and fused with model predictions to stabilize labels and provide interpretable reasoning. In a controlled evaluation, we show that contextual inference improves accuracy, reduces label volatility, and enhances robustness under ambiguous conditions. We also report findings from an eight-week in-home deployment with 14 deaf and hard-of-hearing participants, revealing how context-aware feedback and explanations shape users' trust, understanding, and correction strategies. Our results demonstrate the viability of context-integrated sound recognition in everyday environments.

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

Additional Key Words and Phrases: Sound recognition, Accessibility, Deaf and hard of hearing

ACM Reference Format:

Dhruv Jain and Jason Miller. 2026. SoundTrace: Integrating Temporal Context and Episodic Memory for Real-Time Sound Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 113 (September 2026), 29 pages. <https://doi.org/10.1145/3831645>

1 Introduction

Environmental sound is a subtle but vital thread of everyday awareness: door knocks, appliance beeps, the rhythm of footsteps, the hum that signals life continuing as usual. For deaf and hard-of-hearing (DHH) people, these cues are often inaccessible, and existing assistive technologies only partially bridge that gap. Mainstream sound recognition systems can identify common events [27, 28, 51], but they operate in context-free modes, treating every sound as an isolated occurrence, detached from where it happens, when it happens, or what usually follows. As a result, users encounter alerts that are unpredictable, repetitive, or ambiguous—leaving them to wonder not just *what* happened, but *why now*?

In this work, we explore how sound recognition might become *contextually aware*, *interpretable*, and *adaptive over time*. We introduce *SoundTrace*, a real-time, on-device sound recognition system that integrates temporal context and episodic memory to reason about patterns of everyday activity. Rather than classifying each audio window in isolation, SoundTrace learns from its own experience: it remembers which sounds recur in which places and times, and it uses those memories to make current inferences more stable, meaningful, and explainable. For instance, a beep detected at 7 p.m. in the kitchen is more likely to be a microwave than an alarm clock; a

Authors' Contact Information: Dhruv Jain (*corresponding author), Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA, profdj@umich.edu; Jason Miller, Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA, jmiller@umich.edu.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2026 Copyright held by the owner/author(s).

ACM 2474-9567/2026/9-ART113

<https://doi.org/10.1145/3831645>

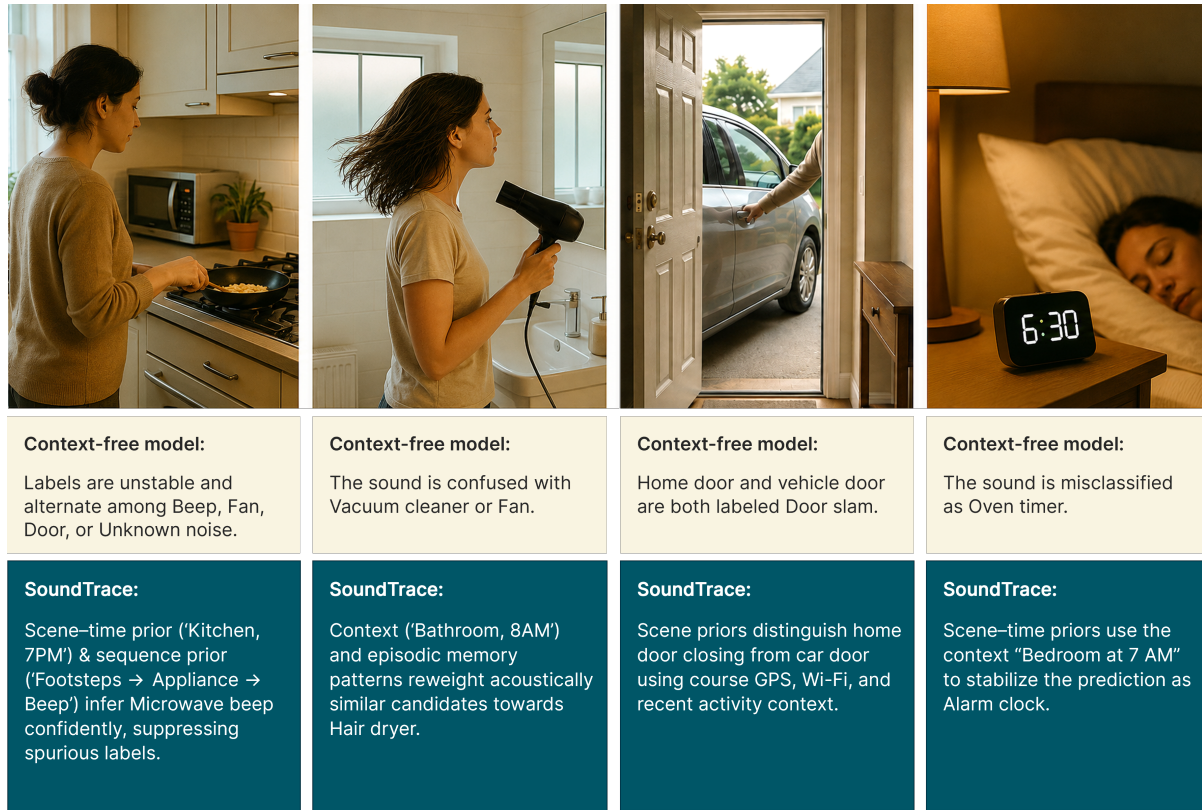


Fig. 1. Four scenarios demonstrating failure cases of context-free sound recognition and how SoundTrace resolves them through contextual inference and memory-based reasoning.

door thud after footsteps is probably someone entering, not furniture shifting. By embedding such temporal and situational regularities directly into the inference loop, SoundTrace grounds recognition in the lived rhythms of daily life.

Technically, SoundTrace augments a neural audio classifier with lightweight memory structures that store symbolic event traces, estimate scene-time and short-range sequence priors, and update those priors through user corrections. During inference, these memory-derived priors are retrieved according to the current scene, time, location cluster, and recent event history, then fused with model predictions to stabilize labels. Natural-language explanations expose the reasoning behind each inference—how model confidence, contextual priors, and sequence cues contribute to the final decision. This design differs from conventional temporal smoothing because memory influences prediction through context-specific evidence accumulated over days and weeks, rather than only enforcing local label continuity. A correction interface lets users refine system behavior in situ, forming a human-in-the-loop learning cycle that adjusts over weeks of use. All computation runs locally on smartphones, ensuring privacy and offline operation.

Our work sits at the intersection of accessibility, machine listening, and human-AI interaction. It extends a long trajectory of DHH sound-awareness research that has evolved from fixed home installations to mobile,

personalizable, and increasingly interactive systems. While prior systems such as *ProtoSound* [25] and *AdaptiveSound* [10] explored personalization through retraining or feedback loops, they largely treated each sound as a fresh classification problem, with limited memory of prior context. Likewise, memory-augmented AI systems in HCI (e.g., MemPal [38] or Mirai [11]) have shown how episodic context can scaffold reasoning, though most remain focused on recall and reflection rather than real-time, longitudinal assistive sensing. SoundTrace unites these threads: a memory-augmented, context-aware, and user-correctable recognizer designed for everyday accessibility.

We evaluated SoundTrace through three complementary studies. A controlled quantitative experiment benchmarked accuracy, stability, and calibration against baseline models under noisy conditions. A laboratory study with 8 DHH participants probed interpretability, comparing explanations and correction interfaces with a non-explainable version. Finally, an eight-week in-home deployment with 14 DHH participants examined how contextual adaptation, explanations, and feedback integrate into domestic routines, shape trust, and reveal new forms of acoustic awareness. Together, these studies connect technical performance with lived experience, illustrating how context and memory reshape how sound recognition is experienced in practice.

In sum, our contributions are:

- (1) A real-time, on-device sound recognition architecture that integrates neural inference with contextual priors derived from symbolic episodic memory, enabling stable, privacy-preserving recognition in everyday environments;
- (2) A memory-augmented inference mechanism that stores, retrieves, and updates scene-time and sequence evidence without retaining raw audio or requiring cloud retraining; and
- (3) Empirical evidence from controlled, lab, and field studies showing that contextual integration improves accuracy, stability, robustness, interpretability, and longitudinal user trust.

More broadly, this work reframes assistive sound recognition as a form of *situated intelligence*—a companion system that learns with its users and grounds perception in memory, transforming sound recognition from a momentary act of classification into an ongoing, situated process of sensemaking.

2 Related Work

We situate our contribution across assistive sound awareness systems, contextual audio modeling, memory-augmented AI, and co-adaptive intelligence that bridges human interpretability with continuous, resource-constrained inference.

2.1 Sound Awareness and Accessibility for DHH Users

Research on sound awareness systems for deaf and hard-of-hearing (DHH) users has evolved from ambient, hub-based devices toward personalized, mobile, and interactive tools. Early in-home systems focused on ambient visualization and shared awareness [21, 39]. *HomeSound* [27] reported an iterative, multi-home deployment of a hub-style device that sensed, visualized, and offered explanations for domestic sounds. The authors highlight design tensions around false positives, glanceability, and the sociality of sound in shared spaces—issues that recur in our own field study (e.g., when and how a system should “break into” attention).

Mobile form factors then became more prominent. Liu et al. [51] built a smartphone-based acoustic event recognition and notification system for DHH students, embedding a lightweight CNN and energy-efficient sensing logic. They conducted evaluations and user studies with 86 participants, showing the feasibility of on-device recognition across varied contexts. *SoundWatch* shifted the paradigm towards wearables, exploring smartwatch-based recognition and feedback [28]. In a three-week field deployment with 10 DHH users [22], SoundWatch exposed real-world constraints—battery life, false alarms, and notification pacing. A later CACM research highlight [29] distilled engineering and UX lessons across architectures (watch-only vs. watch+phone/cloud), reinforcing

the need for low-latency, on-wrist feedback and minimal setup burden—constraints we adopt in SoundTrace’s smartphone-centric design.

More recent work emphasizes personalization and end-user control. Goodman et al. [17] introduced an interface for direct model retraining on smartphones, revealing the tension between personalization gains and user burden from labeling and fatigue. *Spectra* [18] refined this vision through an interactive machine-learning framework that allows users to curate, correct, and blend custom sound profiles, advancing the agenda of personalizable recognition in everyday contexts. Their findings underscore that personalization is not only about accuracy but also about agency—the ability to shape system behavior in situ.

Building on these foundations, *ProtoSound* [25] demonstrated end-user training of sound categories with few examples, enabling on-device customization while surfacing challenges in recording interfaces and error recovery—gaps we address through contextual priors and episodic memory rather than raw-audio retention. *AdaptiveSound* [10] furthered this agenda by integrating error correction feedback loops, showing that user-facing negotiation of AI uncertainty fosters trust and engagement. Recent systems push toward collaborative and orchestrated adaptation: *HACS* [24] framed human–AI co-design workflows for designing sound awareness systems where people and models jointly specify system behavior, while *SoundWeaver* [23] dynamically “weaves” outputs of multiple AI models around a DHH user’s goals and contexts via a heads-up display, illustrating that temporal and situational framing—not just classifier accuracy—shapes perceived usefulness.

Early context-aware audio systems also demonstrated how auditory cues can provide peripheral awareness of information embedded in physical and social contexts. *Audio Aura* [41] used environmental audio cues to deliver background information associated with users’ activities and locations, while *Nomadic Radio* [49] explored wearable audio notifications designed to preserve awareness while reducing interruption burden. These systems established the broader value of contextual auditory awareness, but they primarily focused on delivering information through audio. SoundTrace addresses a different accessibility problem: interpreting environmental sound events for DHH users. In SoundTrace, context is not only an output design consideration but an input to inference itself, shaping how acoustic predictions are stabilized, explained, and adapted over time.

In sum, this body of work surfaces three persistent needs: (1) *contextual fit*: reducing flicker and “why now?” confusion through temporal grounding; (2) *legibility and control*: enabling users to understand and correct system behavior; and (3) *lightweight adaptation*: supporting personalization without heavy data retention or cloud dependency. SoundTrace advances these aims by integrating temporal context and episodic memory to stabilize predictions, expose inference rationales, and allow user corrections to continually shape contextual priors (see Figure 1).

2.2 Context and Temporal Continuity in Sound Recognition

On the technical side, modern environmental sound recognition draws from large-scale supervised and self-supervised audio learning. *PANNs* (CNNs pretrained on AudioSet [13]) established broadly transferable front-ends for tagging and event detection—relevant for small, efficient edge models [30]. Transformer-based models such as *AST* and *SSAST* showed that attention-only spectrogram models can match or exceed CNNs, with self-supervised pretraining reducing labeled-data requirements—important for personalization and long-tail sounds [14, 15]. In parallel, self-supervised speech/audio embeddings like *wav2vec 2.0* and *BEATs* deliver robust representations under noise and domain shift, with *BEATs*’ discrete tokenization particularly relevant to our design choice to store quantized, symbolic traces rather than raw waveforms [3, 4, 7].

Yet most recognition pipelines remain *context-agnostic* at inference time: labels are predicted per window or event with little regard for *where* and *when* the event occurs, or *what just happened*. Temporal smoothers (e.g., HMMs/CRFs) improve continuity but do not encode user- or place-specific regularities. Conversely, acoustic scene classifiers are often trained and run offline, disconnected from streaming recognition. SoundTrace fills

that gap by fusing a *scene-time prior* (based on hour, coarse location / Wi-Fi clusters) with a *sequence prior* over nearby events, combined with model logits. With episodic summarization, this approach balances continuity, locality, and memory constraints on-device.

2.3 Memory-Augmented and Reflective AI Systems

A growing body of work in HCI and AI explores memory augmentation—enabling systems to accumulate, summarize, and recall structured experiences over time to better mediate between raw sensing and user-facing behavior. In many of these designs, memory is not just storage but a *substrate for interpretation, context, and proactive scaffolding*.

One influential inspiration is the *Generative Agents* framework [45], which layers episodic memory, reflection, and planning atop language models to yield temporally coherent behaviors in simulated agents. While generative agents operate in narrative domains, the architectural idea—retrieve relevant prior traces, reflect, and condition future action—resonates with how SoundTrace draws on episodic context to stabilize sound inference over time.

In more applied HCI settings, memory-aware systems are emerging. *Mirai* [11] is a wearable assistant that anticipates user state transitions and delivers voice nudges, leveraging stored context and expected future needs rather than reacting to instantaneous input. Although the evaluation focuses on controlled use cases, *Mirai* illustrates how memory can mediate when and why an AI intervenes. The *Goldilocks Time Window* [12] further theorizes how systems should time interventions optimally and provides guidance on how memory horizons should be shaped for meaningful action.

Projects such as *MemPal* [38] aim to offer wearable memory assistance for everyday life, especially for older adults: they emphasize lightweight, privacy-preserving capture (e.g., feature-level logging instead of raw video), episodic clustering, and context retrieval. While not yet deployed in long-term field studies, *MemPal*'s focus on compressed memory echoes our design choice to retain symbolic event traces rather than full audio or transcripts.

These systems tend to adopt one of two trade-offs: either capture dense, continuous streams (such as raw audio, video, or transcripts) and operate over short, controlled scenarios, or limit memory to simplified logs and restrict interaction complexity. For example, *Memoro* [54] continuously transcribes audio into rich, timestamped text logs; however, its user study takes place in conversational lab settings rather than open homes. Such approaches often face scalability or retention constraints—memory becomes excessive, retrieval slows, or users disengage from reviewing logs.

SoundTrace positions itself between these extremes: we adopt compact, symbolic representations, episodic summarization and decay, and on-device retrieval to enable long-term, streaming use in the real world. Our system shows that memory augmentation need not imply unbounded data capture; rather, with careful compression and fusion, memory can act as a *stabilizing scaffold* for perceptual inference and user reinterpretation across weeks without growth blowup. In doing so, SoundTrace contributes a novel instantiation of memory-augmented interaction tailored to continuous perceptual systems rather than solely narrativ or recall use.

2.4 Explainable, Co-Adaptive, and Situated AI

HCI scholarship consistently shows that transparency and interpretability are more than niceties—they are essential for trust, calibration, and user agency. Lim et al.'s [35] study of “Why/Why-Not” explanations showed that users prefer explanations about *why* a decision was made—not just *why not*—which improves both trust and mental models. This insight motivates our structured natural-language explanations, which attribute outcomes to interpretable factors such as model confidence, scene-time priors, and sequence priors. Kulesza et al.'s *Explanatory Debugging* [31] introduced an iterative cycle where users inspect model reasoning, provide corrections, and observe updated behavior—an interaction mirrored in our interface, where corrections update contextual priors and recalibrate future predictions.

Building on this foundation, recent work in human-centered explainable AI emphasizes that transparency must be coupled with user agency and context sensitivity [33]. Simply exposing internal logic can backfire if users lack the means to act on it, or if explanations ignore situational nuance. In SoundTrace, explanations are explicitly contextual: they reference the current scene, temporal window, and contributing model signals. Users can rename or merge scene clusters, tune fusion weights, and observe how feedback reshapes future inference—transforming transparency from passive disclosure into a form of participatory control.

Finally, drawing from Suchman’s theory of *situated action* [52], we view interpretation as inherently contextual. Systems that ignore the social and material particularities of place and timing risk brittle or inappropriate behavior. Our scene–time priors and episodic memory are a pragmatic instantiation of this lens under resource constraints, grounding inference in the lived rhythms of daily life rather than abstract, context-free sound classification.

2.5 Summary of Gaps and Our Contribution

Across accessibility, audio machine learning, memory-augmented interaction, and explainable AI, a persistent gap remains: real-time systems that reason with *compact memory* about what tends to happen *here, now, and next* while keeping users in the loop. SoundTrace contributes: (1) an episodic, symbolic memory that stores quantized features and context tags; (2) scene–time and sequence priors fused with model logits to stabilize labels and reduce volatility; and (3) user-facing explanations and corrections that update contextual priors entirely on device. Compared with personalization-by-retraining systems, SoundTrace adapts through symbolic contextual priors rather than modifying the underlying acoustic model. Compared with temporal smoothing methods such as HMMs or CRFs, it models both short-range event continuity and longer-term scene–time regularities. Compared with context-aware notification systems, it uses context to directly influence recognition probabilities rather than only determining when or how to notify. Together, these elements position SoundTrace as a memory-augmented inference pipeline for situated sound recognition rather than a new acoustic classifier. In doing so, it connects personalization with temporal continuity, demonstrating that contextual inference can yield both technical improvements and experiential gains for DHH users in everyday environments—complementing and extending systems such as *ProtoSound* [25], *AdaptiveSound* [10], *HACS* [24], and *SoundWeaver* [23].

3 System Design and Implementation

We designed and implemented *SoundTrace*, a lightweight, memory-augmented environmental sound recognition system that augments neural audio classifiers with context-sensitive reasoning and user-facing interaction mechanisms. This system was developed to investigate how long- and short-term context—derived from local observations of environmental audio—can enhance the usability, explainability, and adaptability of real-time sound recognition. We describe the design objectives, system architecture, representative example scenarios, and technical implementation that supports real-time, context-aware operation.

3.1 Motivation and Design Objectives

Environmental sound recognition has advanced rapidly with deep learning approaches such as convolutional networks (e.g., PANNs [30]), audio transformers (e.g., AST [14]), and self-supervised audio embeddings (e.g., wav2vec 2.0 [3], BEATs [7]). These methods achieve strong performance on isolated sound events but typically operate in context-free conditions, assuming that events are independent and identically distributed. This diverges from how humans perceive everyday acoustic scenes, where meaning depends on what happens before, where, and when a sound occurs.

SoundTrace bridges this gap by introducing a lightweight, on-device model that combines neural inference with contextual reasoning and episodic memory. The design goal was to achieve real-time, explainable, and adaptive operation without raw audio storage or cloud processing. This entailed several core objectives:

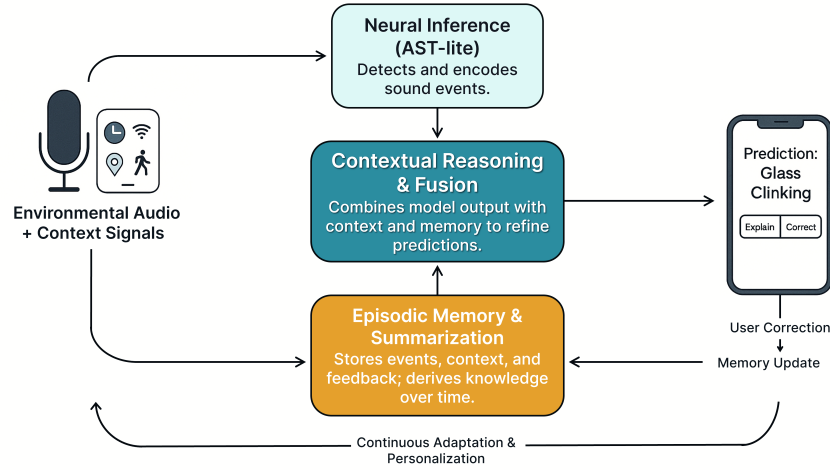


Fig. 2. **SoundTrace System Pipeline.** The system processes both acoustic and contextual information to generate adaptive, interpretable sound recognition results. From left to right: environmental audio and on-device context signals flow into the *Neural Inference (AST-lite)* module, which detects and encodes sound events. Embeddings and metadata are stored in the *Episodic Memory & Summarization* module, which accumulates experience and derives contextual knowledge over time. The *Contextual Reasoning & Fusion* layer integrates neural predictions with context and memory-derived insights to refine output probabilities. The final predictions are shown in the *User Interface*, where users can view explanations or provide corrections. Corrections update the memory and contextual priors, creating a continuous loop of *adaptation and personalization*.

- (1) **Real-time performance:** Process and respond to sounds with sub-second latency on mobile devices.
- (2) **Memory-efficient operation:** Retain symbolic and quantized features rather than raw audio.
- (3) **Contextual reasoning:** Use past observations to inform present inferences.
- (4) **Explainability and user trust:** Provide clear rationales for classifications and enable user correction.
- (5) **Continuous adaptation:** Incrementally adapt to changing acoustic environments over time.
- (6) **Privacy by design:** Operate entirely on-device to minimize data exposure.

These objectives balance technical feasibility with usability-centered values essential for accessibility technologies and everyday home or workplace deployments.

3.2 Overall Architecture

The SoundTrace architecture comprises three interdependent layers (Figure 2):

- An **acoustic frontend** that detects and classifies sound events in real time;
- A **contextual reasoning layer** that augments neural predictions with memory-derived priors; and
- A **long-term episodic memory** that supports adaptation, explanation, and summarization.

Audio features and on-device context signals flow through the pipeline to produce a fused, interpretable prediction that the user can review and correct. User feedback and new observations continuously update the episodic memory, closing the loop for incremental learning.

3.3 Acoustic Frontend: Embeddings and Event Tokens

The acoustic frontend processes the continuous microphone stream into overlapping 1.0 s windows (0.5 s hop). Each segment is transformed into a 64-band log-mel spectrogram, standardized, and fed to an Audio Spectrogram Transformer (*AST-lite*)—a reduced variant of the full AST architecture [14] designed for CPU efficiency.

AST-lite Architecture. The original AST [14] employs a ViT-Base transformer backbone with 12 attention layers, 768-dimensional embeddings, and 16×16 spectrogram patches. To make inference feasible on mobile CPUs, *AST-lite* reduces this to 6 transformer encoder blocks with 8-head self-attention and 384-dimensional embeddings. Frequency patch resolution is pruned to 8×8, cutting both parameters and compute by roughly 60%. The model is pretrained on AudioSet [13] and fine-tuned on ESC-50 [46] and UrbanSound8K [48] to generalize across household, urban, and ambient environments. The resulting model supports 35 everyday sound categories, including appliance alerts (e.g., microwave, kettle, oven timer), human activity (e.g., footsteps, speech-in-mix, door close), and environmental sounds (e.g., rain, traffic, wind, TV audio). Before finalizing, the class set was refined to emphasize sound categories reported as meaningful to DHH people in prior work [5, 26].

Training formulation and labeling. *AST-lite* is trained as a multi-label audio-tagging classifier, reflecting the polyphonic nature of real-world environmental audio where multiple sound sources may co-occur within a single segment. Training data from AudioSet, ESC-50, and UrbanSound8K are unified into the 35-class target label space. For datasets originally annotated with single labels, labels are represented as one-hot vectors; for AudioSet, available multi-label annotations are retained. The model is optimized using binary cross-entropy over all classes rather than softmax cross-entropy, allowing multiple labels to receive non-zero probability within the same window. At inference time, SoundTrace retains the top- k candidate labels and scores for contextual fusion. For evaluation, Top-1 and Top-3 metrics are computed from this ranked output.

For each 1.0 s segment, *AST-lite* outputs:

- A vector of class probabilities (*logits*) representing candidate sound categories, and
- A dense audio embedding $\mathbf{e} \in \mathbb{R}^{384}$ capturing high-level acoustic features.

Voice Activity Detection and Event Merging. SoundTrace applies VAD before *AST-lite* inference to suppress silence and low-energy regions and reduce unnecessary computation. Only windows that pass this gate are processed by the classifier. Adjacent classified windows with similar top-1 class and high embedding similarity (cosine distance $\delta < 0.15$) are then merged using a class-aware event merger, forming single *event tokens*. This ordering ensures that VAD acts as an acoustic activity gate, while event merging operates after classification to form temporally coherent units for contextual reasoning. Each token captures a temporally consistent sound episode, storing:

- Start and end timestamps,
- Top- k class probabilities and scores,
- Quantized embeddings (8-bit integer, symmetric post-training quantization), and
- Associated on-device context metadata (Section §3.4).

This tokenization ensures that transient predictions are consolidated into meaningful, interpretable sound units.

3.4 On-Device Context Signals

In parallel with acoustic inference, SoundTrace continuously samples lightweight device signals available from the smartphone. These include:

- **Time of day:** Hour-level temporal granularity;
- **Location context:** GPS coordinates rounded to ~100 m grids;
- **Wi-Fi SSID:** Used as a scene proxy when indoors;
- **Motion state:** Stationary, walking, or driving inferred from accelerometer data; and

- **Ambient energy:** RMS energy of the incoming audio stream.

These signals are fused to infer situational context such as “Home–Evening–Stationary.” SoundTrace maintains a dynamic *scene dictionary* by clustering GPS and Wi-Fi traces over time using K-means, capturing frequently visited locations (e.g., *Home*, *Work*, *Campus*) (Figure 4.4). SoundTrace does not infer room-level location from GPS alone. GPS coordinates rounded to approximately 100 m grids are used only for coarse place separation, such as home, campus, or neighborhood. Indoors, finer scene labels are treated as user-interpretable context clusters rather than metrically precise room estimates. These clusters are derived from repeated co-occurrences among Wi-Fi context, temporal patterns, acoustic event distributions, motion state, and user edits. For example, a “Kitchen–Evening” context may emerge because appliance sounds, stationary phone placement, home Wi-Fi, and evening timing repeatedly co-occur; it is not estimated from GPS precision. Users can rename, merge, or delete clusters via the app interface for interpretability and control, making the scene dictionary a controllable interpretive layer rather than an automatic indoor-localization claim. The resulting contextual metadata accompanies every event token and is later used for memory storage and reasoning.

3.5 Episodic Memory and Summarization

SoundTrace operationalizes episodic memory as an active inference substrate rather than a passive event history. Each detected sound is converted into a symbolic event token containing its timestamp, top- k acoustic labels and scores, quantized embedding, inferred or user-edited scene label, Wi-Fi/GPS-derived context, motion state, ambient energy, and any user correction. Raw audio is discarded after inference. The stored representation is therefore compact, symbolic, and privacy-preserving.

To support long-term reasoning and adaptation, SoundTrace organizes event tokens into *episodes*—contiguous sequences of related activity typically spanning 5–30 minutes. Episode boundaries are detected by analyzing energy shifts, embedding drift (cosine distance >0.25), and idle gaps—periods of acoustic inactivity longer than 45–60 s, indicating transitions between activities. For example, a cluster of “Microwave End → Footsteps → Door Close” may form one episode, while a long silence before “TV Audio” begins marks a new episode boundary. Each episode is assigned a unique ID and stored in a persistent SQLite database as a symbolic, privacy-preserving record of experience.

Each stored token includes:

- Timestamps (start and end),
- Top- k predicted labels and scores,
- Quantized embedding,
- Scene label (user-assigned or inferred),
- GPS and Wi-Fi context,
- Motion state and ambient energy, and
- User corrections (if any).

During inference, the current event token retrieves memory along three dimensions. First, SoundTrace retrieves scene–time statistics for the current scene label, hour bucket, and location cluster. Second, it retrieves recent sequence statistics conditioned on the last one or two recognized events. Third, when available, it retrieves nearest exemplar embeddings within the current scene to support disambiguation among acoustically similar labels already proposed by the acoustic model. Memory therefore does not introduce arbitrary labels; instead, it reweights plausible acoustic candidates according to prior events observed in similar contexts.

Episodes sharing consistent context (scene, motion, or location) are grouped by spatiotemporal continuity into coarse situational contexts (e.g., “Home–Evening–Stationary” or “Commute–Morning–Moving”; see Figure 3 for an example).

Periodic summarization then condenses memory through:

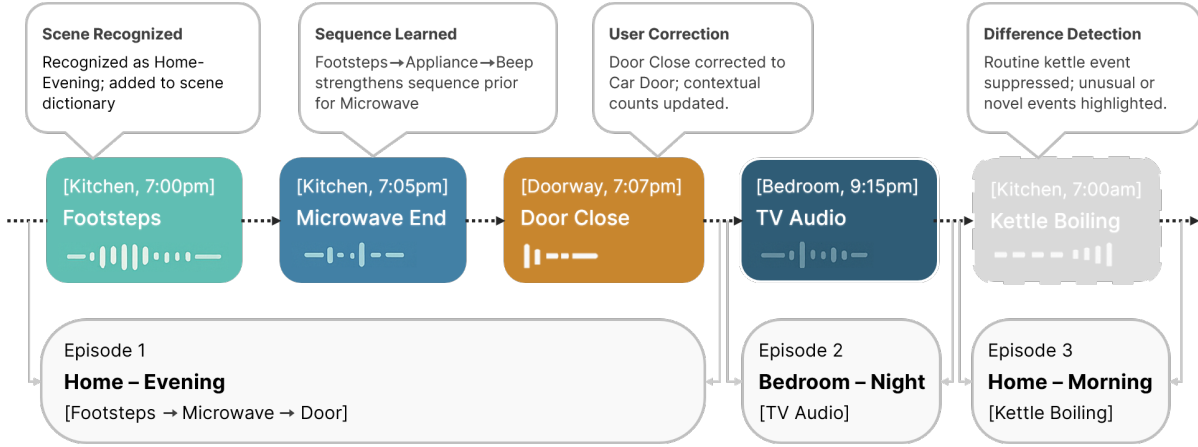


Fig. 3. **Episodic Memory Trace.** How individual sound events evolve into structured episodic context through reasoning and user interaction. Five events—[Kitchen, 7:00 pm] *Footsteps*, [Kitchen, 7:05 pm] *Microwave End*, [Doorway, 7:07 pm] *Door Close*, [Bedroom, 9:15 pm] *TV Audio*, and [Kitchen, 7:00 am] *Kettle Boiling*—form three contextual episodes (*Home–Evening*, *Bedroom–Night*, and *Home–Morning*). Colored callouts highlight how information is processed: (a) **Scene recognized**—the system links detected sounds to a scene–time cluster (e.g., “Home–Evening”); (b) **Sequence learned**—patterns such as *Footsteps* → *Appliance* → *Beep* reinforce event relations; (c) **User correction applied**—edits such as changing “Door Close” → “Car Door” update contextual counts; and (d) **Difference detection**—repeated daily events (e.g., kettle) are down-weighted while novel ones are prioritized.

- Clustering event embeddings within episodes to merge redundant tokens;
- Selecting exemplars for vector search and nearest-neighbor retrieval;
- Computing frequency statistics (dominant labels, frequent transitions, and peak times);
- Pruning non-exemplars after a retention window (e.g., 7 days) to limit growth.

This process is analogous to memory consolidation in cognitive science [40], converting short-term observations into semantic summaries.

3.5.1 Deriving Contextual Priors. From these episodic summaries, SoundTrace derives two probabilistic priors that capture regularities in user environments:

- **Scene–Time Prior:** The frequency of each sound y given a scene s , hour h , and location cluster g :

$$P_{\text{scene}}(y \mid s, h, g) = \frac{C(y, s, h, g) + \lambda}{\sum_{y'} C(y', s, h, g) + \lambda K} \quad (1)$$

- **Sequence Prior:** The probability of an event y_t following a sequence of preceding events:

$$P_{\text{seq}}(y_t \mid y_{t-1}, y_{t-2}) = \frac{C(y_{t-2}, y_{t-1}, y_t)}{\sum_{y'} C(y_{t-2}, y_{t-1}, y')} \quad (2)$$

Here, $C(\cdot)$ denotes frequency counts within episodic memory, λ is a Dirichlet prior (e.g., $\lambda = 1$), and K is the number of labels. Both priors are updated incrementally with exponential decay ($\rho = 0.98$ per day), ensuring that older contexts gradually lose influence as environments change. User corrections are incorporated as higher-confidence memory updates. When a user changes a predicted label y to y' , SoundTrace increments the corrected label within the relevant scene–time and sequence contexts and downweights the incorrect label in those same contexts. These updates affect subsequent predictions without retraining the acoustic model. Exponential decay

reduces the influence of stale routines, allowing the system to adapt when household patterns, device placement, or user behavior changes. These priors are supplied to the contextual fusion module during inference.

3.6 Contextual Fusion and Reasoning

During inference, SoundTrace fuses three complementary sources of evidence:

- (1) Neural predictions from the acoustic model;
- (2) Contextual priors derived from memory (Scene–Time and Sequence); and
- (3) Real-time context signals (scene label, time, motion).

The fused posterior probability of each label y is computed as:

$$\log P^*(y) = \alpha \log P_{\text{model}}(y) + \beta \log P_{\text{scene}}(y) + \gamma \log P_{\text{seq}}(y) \quad (3)$$

where $\alpha + \beta + \gamma = 1$ (default $\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$).

Temporal smoothing. To reduce flicker and ensure stability, SoundTrace applies a streaming Hidden Markov Model (HMM) smoother [47] over the last $N=5$ tokens, using transition probabilities initialized from the sequence prior. Unlike offline Viterbi decoding, the smoother performs forward-only updates in constant time, maintaining low latency while enforcing temporal coherence. This produces stable, interpretable event predictions displayed to users through the interface. User corrections and contextual updates then propagate back into episodic memory, adjusting priors for future inference (Figure 2).

The fusion layer separates three mechanisms that operate at different temporal scales. The acoustic model captures instantaneous spectral evidence. The scene–time prior captures longer-term regularities, such as which sounds tend to occur in a given context and hour. The sequence prior captures short-range event structure, such as sounds that commonly follow footsteps, door activity, or appliance use. The HMM smoother is applied only after this fusion step to reduce display-level flicker. Thus, contextual memory contributes directly to the posterior before smoothing, distinguishing SoundTrace from approaches that only post-process labels for temporal continuity.

3.7 Interaction, Explanation, and Correction

Each prediction in SoundTrace is announced through a brief vibration pulse and accompanied by a natural-language explanation that reveals how model and contextual priors contributed to the fused probability. Displayed beneath the predicted label in the *Explanation & Correction* screen (Figure 4.2), the text summarizes the weights of the acoustic model, scene–time prior, and sequence prior. For example:

“Prediction: *Microwave Beep* (0.82). Model 0.63 + Scene 0.13 + Sequence 0.06.”

Here, 0.82 denotes the fused probability, while the additive terms indicate contributions from the contextual priors.

Explanations are produced by a lightweight, rule-based text composer that transforms the fusion layer’s numerical outputs into short, human-readable sentences. When contextual priors dominate, the phrasing shifts toward a more descriptive form, such as:

“Prediction: *Oven Timer*—common in the *Kitchen* at 7 PM, increasing confidence.”

To balance clarity and flexibility, SoundTrace uses a two-tier explanation model. A concise one-sentence summary is shown by default, while a longer contextual rationale can be accessed through a long-press gesture on any prediction. The expanded view includes information about acoustic evidence, scene label, time of occurrence, and preceding events (e.g., “Footsteps → Appliance → Beep”).

These explanations are accompanied by user correction options (*Correct*, *Edit*). When a user corrects a label, SoundTrace updates contextual priors by upweighting the confirmed label and downweighting the incorrect one within the relevant scene–time and sequence contexts. These corrections are logged as ground truth for subsequent adaptation.

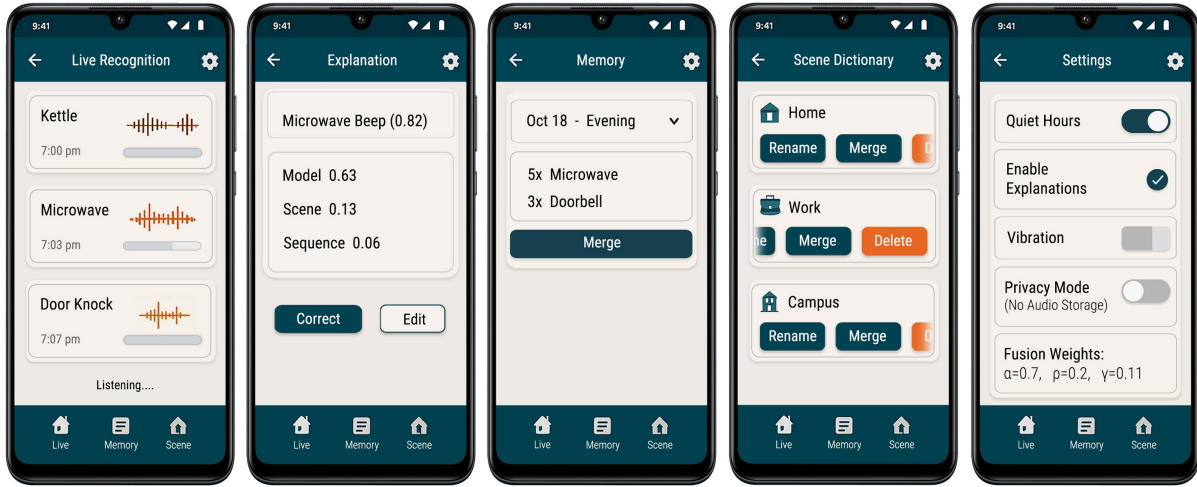


Fig. 4. **SoundTrace App Interface.** The mobile application comprises five coordinated screens supporting interpretability, correction, and contextual control. **1. Live Recognition.** Displays real-time sound detections as a scrollable list with event names, timestamps, and confidence bars. Users can tap any entry to open its explanation view. **2. Explanation & Correction.** Presents the system’s reasoning for a selected event, decomposing the fused probability into model, scene, and sequence contributions. Quick actions (*Correct*, *Edit*) allow users to refine labels, immediately updating contextual priors. **3. Memory / History.** Shows a timeline of past episodes grouped by scene and time (e.g., *Kitchen—Evening*). Episode cards summarize frequent sounds and enable retrospective browsing or filtering by label and location. **4. Scene Dictionary.** Lists automatically inferred context clusters (e.g., *Home*, *Work*, *Campus*) with options to *Rename*, *Merge*, or *Delete*. Edits propagate to episodic memory and update scene–time and sequence priors. **5. Settings.** Provides controls for quiet hours, explanation verbosity, vibration intensity, privacy mode, and fusion weights (α, β, γ).

3.8 Example Scenarios and Differentiation

To illustrate how SoundTrace improves on existing sound recognition systems such as *SoundWatch* [28], we describe six everyday scenarios that represent common challenges in environmental sound awareness. Each example contrasts the existing context-free systems with how SoundTrace uses context and memory to produce more stable and useful results. Figure 1 visualizes four of these examples.

The following scenarios illustrate contexts in which memory can improve recognition, but not every recognized sound is equally informative in every situation. When the user directly produces a sound, such as operating a hair dryer or vacuum cleaner, recognition may provide limited value for identification because the user already knows the source. In such cases, SoundTrace may still support confirmation, logging, appliance-state awareness, or retrospective review, but its primary accessibility value lies in ambient awareness of externally generated or missed sounds, such as door activity, alarms, appliances completing cycles, or actions by other people in shared spaces. Contextual reasoning is therefore most useful when the sound source is outside the user’s visual field, partially attended, overlapping with other sounds, or ambiguous from acoustic evidence alone.

3.8.1 Temporal Fragmentation: Overlapping Sounds in Everyday Spaces.

Challenge: Many environmental sounds occur together or in quick succession, confusing models that process each window independently.

Example: In a kitchen, the microwave beeps while the stove hood runs and footsteps pass by.

Existing systems: Alternate unpredictably between labels such as “*Beep*,” “*Fan*,” or “*Door*,” producing flickering,

unstable predictions.

SoundTrace: Combines short-term events through a *sequence prior* (“Footsteps → Appliance → Beep”) and *scene–time context* (“Kitchen, 7PM”) to infer *Microwave Beep* confidently while suppressing spurious transitions.

3.8.2 Spectral Ambiguity: Similar Sounds, Different Meanings.

Challenge: Acoustically similar appliances or devices are easily confused when models rely only on raw audio features.

Example: A hair dryer-like sound is detected near the bathroom while the user is elsewhere in the home, or while another household member is using an appliance.

Existing systems: Misclassify such sounds as *Vacuum* or *Fan*, ignoring situational cues.

SoundTrace: Integrates *scene and time priors* (“Bathroom, 8AM”) with *episodic memory* of past events to distinguish likely bathroom activity from acoustically similar vacuum or fan sounds. When the user is the direct source of the sound, the same recognition may be more useful for logging or confirmation than for immediate identification.

3.8.3 Scene Confusion: Same Sound, Different Place.

Challenge: Identical acoustic patterns can have different meanings depending on location.

Example: The same door-closing sound occurs indoors and outside near a car.

Existing systems: Label both as *Door Slam*, missing the environmental difference.

SoundTrace: Uses *scene recognition* derived from Wi-Fi and GPS clusters to differentiate *Indoor Door Closing* from *Car Door Closing*, producing location-specific interpretations.

3.8.4 Repetitive Misclassification: Repeated Sounds Across Time.

Challenge: Daily sounds often repeat but have distinct meanings depending on time and setting.

Example: A morning alarm and an evening oven timer share the same acoustic signature.

Existing systems: Confuse both as *Alarm Clock*, lacking awareness of temporal context.

SoundTrace: Builds *episodic memory* that differentiates by *scene and time*—“Bedroom, 7AM → Alarm Clock” vs. “Kitchen, 8PM → Oven Timer”—learning distinctions from long-term usage patterns. Although users often know the time of day, incorporating temporal context into inference reduces the burden of resolving ambiguous labels manually. SoundTrace uses time and scene not to replace user knowledge, but to make the system’s prediction more stable, interpretable, and less prone to repeated misclassification.

3.8.5 Source Attribution: Who (or What) Is Speaking.

Challenge: Speech-like sounds can come from people, televisions, or devices, yet sound identical to a classifier.

Example: A TV show playing in the living room produces speech that resembles a conversation.

Existing systems: Alternate between *Speech* and *Conversation* labels, showing high volatility.

SoundTrace: Associates *Speech + Evening + Living Room + Stationary Device* with *TV Audio*, using episodic memory to infer the most likely source and stabilize results.

3.8.6 Difference Detection: Knowing What’s New or Important.

Challenge: Traditional recognizers repeatedly notify users about familiar background events, creating fatigue or distrust.

Example: Daily kettle or fridge sounds trigger identical alerts every day.

Existing systems: Treat every sound as new and equally important.

SoundTrace: Performs *difference detection* through episodic summarization, filtering out routine or expected events and surfacing only novel or significant ones (e.g., a siren or a new alarm).

3.9 Implementation and Performance

SoundTrace is implemented as a fully on-device pipeline that integrates neural inference, contextual reasoning, and episodic memory management. The system uses PyTorch Mobile for model inference, torchaudio for preprocessing, SQLite for event storage, and FAISS for embedding indexing. All computation executes locally; no raw audio or embeddings are transmitted off-device.

End-to-end processing (from audio capture to contextualized prediction) completes in under 150 ms per 0.5 s hop on a mid-range Android device (e.g., Pixel 6). The pipeline consumes less than 400 MB RAM and under 40% CPU utilization, supporting continuous background use.

To support long-term operation, SoundTrace employs a bounded episodic memory design. Rather than storing raw audio, the system retains compact symbolic event tokens and quantized embeddings, which are periodically consolidated through a daily summarization process scheduled during idle periods via Android's WorkManager. This process clusters redundant tokens within episodes, retains representative exemplars for retrieval, and prunes non-exemplar entries after a retention window, ensuring that memory growth remains sublinear with respect to the underlying audio stream.

Across the eight-week deployment, the median database size increased from 10.1 MB in Week 1 to 45.7 MB in Week 8, reflecting gradual accumulation of symbolic traces rather than high-bandwidth sensory data. Retrieval operates over the retained exemplar index rather than the full event history, resulting in median retrieval latency below 20 ms. Daily summarization required 7.2 s on average and reduced retained tokens by approximately 75% relative to the uncompressed event stream. Together, these results demonstrate that episodic memory remains compact, efficient, and scalable for continuous, fully on-device operation in everyday settings.

Overall, SoundTrace delivers stable, interpretable, and adaptive sound recognition by coupling deep audio embeddings with contextual reasoning and human-in-the-loop feedback. For replicability and future improvements, the entire system code is open sourced: [link withdrawn for anonymous review].

4 Evaluation

We conducted a comprehensive, mixed-method evaluation of SoundTrace across three complementary studies. First, a controlled *quantitative evaluation* assessed the technical efficacy of contextual inference under streaming constraints and acoustic perturbations (§4.1). Second, a *lab study* with eight DHH participants examined how natural-language explanations and one-tap corrections shape understanding and trust among end users (§4.2). Third, an *eight-week in-home deployment* with 14 DHH participants investigated longitudinal adaptation, everyday use, and lived experience (§4.3). Together, these studies triangulate algorithmic performance with human-centered outcomes, evaluating *accuracy*, *stability*, *robustness*, *interpretability*, and *acceptance* in the wild.

4.1 Controlled Evaluation: Accuracy, Stability, and Robustness

4.1.1 Goals. We evaluated whether contextual inference improves (i) **accuracy** on everyday sounds, (ii) **stability** by reducing short-horizon label volatility, and (iii) **robustness** under noisy and reverberant acoustic conditions (e.g., low SNR, overlapping sources). We hypothesized that the fused model would outperform the base classifier on all three dimensions.

4.1.2 Datasets and Setup. We constructed a multi-context evaluation corpus by combining three sources: ESC-50 [46], UrbanSound8K [48], and a field-collected everyday sound dataset (4.5 hours) adapted from two prior works on sound recognition for DHH people [27, 28]. The latter comprises recordings from diverse real-world settings—including homes, campuses, transit stations, and outdoor streets—captured on mid-range Android devices under varied acoustic conditions (quiet, conversational, and noisy).

We simulated real-time streaming inference by segmenting audio into 1.0 s windows with 0.5 s hop and executing the full SoundTrace pipeline on-device. All datasets were normalized to a consistent 16 kHz sample

rate and annotated using the same 35 target classes supported by our model, spanning both environmental and human activity sounds. We compared four system variants with the same acoustic model (AST-lite) to isolate the effect of incorporating context:

- (1) **Baseline**: AST-lite classifier only (no context, current state-of-the-art – e.g., SoundWatch [28]);
- (2) **+Scene**: Baseline + scene-time prior;
- (3) **+Seq**: Baseline + sequence prior;
- (4) **Full (SoundTrace)**: Baseline + scene-time + sequence priors.

All models ran on-device on a Pixel 6 smartphone (ARM Cortex cores, 6 GB RAM). We used five-fold cross-validation stratified by recording location to assess generalization across contexts.

4.1.3 Metrics. We report macro-averaged **Top-1** and **Top-3** accuracy and **F1**.

Volatility quantifies short-term instability in predicted labels over time. Given a sequence of frame-level labels $y_{1:T}$, volatility is:

$$\text{Volatility} = \frac{1}{T-1} \sum_{t=2}^T \mathbb{I}[y_t \neq y_{t-1}]$$

computed within detected events and averaged across the stream; lower is better. We also report calibration via *expected calibration error* (ECE) with 10 bins.

Robustness is evaluated by adding non-stationary ambient noise (café, traffic) at SNR {0, -5, -10 dB} and applying light room impulse responses (small/medium rooms).

4.1.4 Backbone Generality. To assess the generality of these findings beyond a single acoustic model, we additionally evaluate contextual fusion across multiple backbones: (1) a lightweight CNN classifier, representative of models used in prior HCI sound recognition systems (e.g., SoundWatch [28], HomeSound[27]); (2) AST-lite, a compact transformer-based model; and (3) a BEATs-lite embedding model with a lightweight classifier, representing modern self-supervised audio representations in machine learning. All backbones were evaluated under the same streaming setup, datasets, and evaluation protocol, with identical contextual priors and fusion parameters.

4.1.5 Results. Figures 5 and 6 visualize the quantitative performance of SoundTrace across model variants. The fused model, **Full (SoundTrace)**, outperformed the Baseline on all metrics. As shown in Figure 5 (left), top-1 accuracy improved by 6.5% absolute over the Baseline (79.3 vs. 72.8), macro-F1 by 6.7% (78.1 vs. 71.4), and volatility decreased by 34% (0.22 vs. 0.32). Expected Calibration Error (ECE) also improved from ~0.12 (Baseline) to ~0.08 (Full), indicating better-calibrated probabilities (Figure 5, right). Under severe noise conditions (-10 dB SNR), the Full model retained 84.9% of its clean macro-F1 (66.3 vs. 78.1), compared to the Baseline's 68.3% (48.8 vs. 71.4), demonstrating stronger robustness in degraded acoustic conditions (Figure 6).

Paired *t*-tests across cross-validation folds showed significant gains for Full over Baseline on Accuracy, F1, and Volatility ($p < .01$); effect sizes were medium to large (Cohen's $d = 0.64$ – 0.92). Ablations revealed additive benefits: **+Scene** stabilized predictions in stationary contexts (e.g., appliances), whereas **+Seq** improved temporal continuity in short activity chains (e.g., footsteps → door).

These results indicate that the two contextual priors contribute complementary benefits. The scene-time prior produced the largest gains in stationary domestic contexts, where stable routines made contextual base rates informative. The sequence prior produced the largest reductions in short-horizon label switching, especially around event boundaries and overlapping activities. The full model performed best because it combines these long- and short-horizon constraints.

Consistent with the above findings, contextual fusion improved performance across all backbones. For example, macro-F1 increased from 68.2 to 73.0 for CNN, 71.4 to 78.1 for AST-lite, and 74.8 to 79.6 for BEATs-lite, while

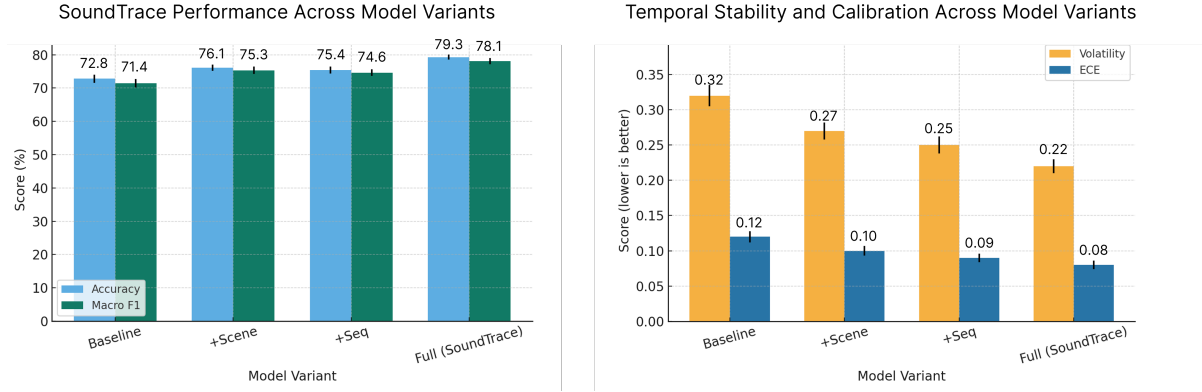


Fig. 5. **Quantitative performance across model variants.** Left: Accuracy and Macro F1 improve from the Baseline through +Scene and +Seq to the Full (SoundTrace) model. Right: Volatility and ECE (lower is better) decline monotonically across the same variants, indicating greater temporal stability and better calibration. Error bars denote one standard deviation across five cross-validation folds.

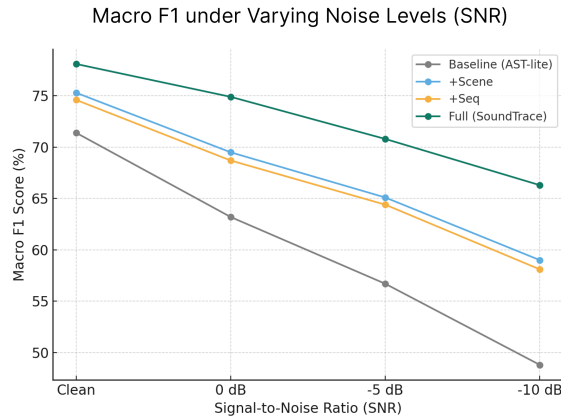


Fig. 6. **Robustness under additive noise (Macro F1 vs. SNR).** Macro F1 for each model variant across noise levels (Clean, 0 dB, -5 dB, -10 dB). The Full (SoundTrace) model sustains the highest F1 and the most gradual decline with increasing noise. Error bars indicate one standard deviation across five cross-validation folds. The numeric values are shown in Appendix A.

volatility decreased by 20–35% across models. Full results appear in Appendix A. While absolute performance varied across architectures, the relative gains from contextual inference were consistent, indicating that SoundTrace’s advantage arises from its memory-augmented inference mechanism rather than any particular acoustic model.

Latency met deployment constraints (< 150 ms per 0.5 s hop), and resource use remained within smartphone limits (CPU < 40%, memory < 400 MB), consistent with Section §3.9.

4.1.6 Takeaways. Contextual priors offered complementary benefits: scene–time priors improved base-rate alignment and probabilistic calibration, grounding predictions in habitual activity patterns; sequence priors, in turn, reduced label flicker and boundary instability by capturing short event chains. Together, the fused

model (SoundTrace) achieved higher accuracy and perceptual smoothness, yielding more stable and trustworthy inferences, especially under noisy conditions. Beyond aggregate improvements, the ablation results clarify why SoundTrace outperforms context-free models. Scene-time priors address global contextual ambiguity by aligning predictions with long-term regularities in specific environments, such as appliance sounds that recur in kitchens or entryway sounds that follow movement. Sequence priors address local temporal instability by discouraging implausible rapid label switching and reinforcing coherent event transitions. These mechanisms operate at different temporal scales and address distinct failure modes. The full model benefits from their combination, demonstrating that SoundTrace's gains arise from explicit modeling of environmental regularities encoded in episodic memory rather than from generic smoothing alone.

4.2 Lab Study: Explanations and Corrections in Use (N=8)

4.2.1 Goals. We investigated how SoundTrace's user-facing features—natural language explanations and one-tap corrections—shape users' understanding, trust, and willingness to engage in in-the-moment adaptation.

4.2.2 Participants. We recruited 8 deaf and hard-of-hearing (DHH) participants (5 F / 3 M; ages 21–58, $M = 37.6$, $SD = 11.2$) through regional Deaf centers, community mailing lists, and social media outreach. Four participants identified as hard-of-hearing, two as Deaf, and one as deaf (small d). Six reported profound hearing loss and two severe. Four participants used hearing aids, two used cochlear implants, and two did not use any hearing assistive devices. Device platforms were balanced (four Android, four iPhone). Six participants had prior experience with sound recognition apps (e.g., Apple Sound Recognition, SoundWatch). Participants received \$50 as compensation.

4.2.3 Procedure. Sessions were conducted in a quiet lab configured to approximate three everyday environments: a *kitchen*, *corridor*, and *office*. Each participant experienced two within-subjects conditions in counterbalanced order:

- (1) **Baseline:** the app displayed only sound labels and confidence values (no explanations);
- (2) **SoundTrace:** the app displayed predictions accompanied by explanations and one-tap correction UI.

After a short tutorial and practice round, participants completed three task blocks (8 clips per block; 24 total). Each block contained short environmental recordings of everyday sounds (e.g., appliance beeps, running water, door activity, speech-in-mix). Audio was streamed through the app to simulate real-time feedback, and participants verbally identified or confirmed the detected sound.

After each block, participants completed two 5-point Likert scales assessing *trust* ("I trusted the system's predictions") and *understanding* ("I understood why it made those predictions"), followed by brief verbal reflections. At the end, a semi-structured interview was conducted (15–20 min), focusing on perceived clarity of explanations, confidence in results, and ease of correction.

4.2.4 Data Analysis. We collected interaction logs, Likert responses, and interview transcripts. Two researchers conducted *Applied Thematic Analysis* [19]. An initial codebook was developed from three transcripts, refined iteratively, and applied to all eight. Inter-rater reliability across the final codebook achieved Cohen's $\kappa=0.81$ and raw agreement of 90.5%. Disagreements were resolved through discussion, and themes were refined through axial coding. We retained negative cases to avoid confirmation bias. Table 1 reports theme prevalence at two levels: participant-level prevalence (the number of participants with at least one coded excerpt per theme) and excerpt-level prevalence (the number and percentage of coded excerpts).

4.2.5 Findings. We summarize salient themes below; quotes are drawn from transcripts but lightly edited for grammar. Table 1 lists theme prevalence across participants and coded excerpts.

Explanations scaffold mental models. Participants reported that linking predictions to time/scene and recent events clarified *why a sound was recognized at that moment*, marking a qualitative shift from guessing

Table 1. Theme prevalence in the lab study (N=8). Counts are descriptive rather than inferential and reflect excerpts coded under the five primary themes; percentages are calculated over these excerpts and do not represent the entirety of interview content.

Theme	Participants	Coded excerpts	% of excerpts
Explanations scaffold mental models	7/8	32	30%
Correction as low-effort agency	8/8	29	28%
Temporal context as tiebreaker	6/8	18	17%
Transparent uncertainty supports trust	6/8	15	14%
Early friction, rapid adaptation	5/8	11	10%

to reasoning. As P4 explained, “*Kitchen, 7:15pm, after footsteps’—that tells me it’s the oven, not my roommate slamming something.*” Ratings of perceived understanding increased from $M=3.1$ (No-Explanation) to $M=4.4$ (SoundTrace), $p < .01$.

Correction as low-effort agency. The one-tap correction afforded quick control without derailing the task. P6: “*If it’s wrong, I tap ‘microwave’ and I’m done. It learns and I move on.*” Participants preferred lightweight edits over free-text. Trust ratings rose from $M=3.2$ to $M=4.1$ ($p < .05$).

Temporal context as tiebreaker. Where acoustic evidence was ambiguous, recent-sequence cues were persuasive. P2: “*It could be door or cupboard—the ‘footsteps then door’ made door more convincing.*” Participants highlighted that time-of-day cues matched expectations for shared routines.

Transparent uncertainty supports trust. Showing confidence and contextual rationale supported calibrated trust: “*If it says 0.62 and explains the pattern, I can decide if that’s good enough.*” (P2).

Early friction, rapid adaptation. Corrections were highest in Block 1 and declined across blocks as priors adapted, consistent with the quantitative stability gains. P5: “*First few were off, then it settled into my kitchen sounds.*”

4.2.6 Summary. Explanations and corrections transformed the interaction from post-hoc judgment to *guided sensemaking*. Participants perceived SoundTrace as teachable rather than opaque, a prerequisite for sustained use. They valued explicit ties to *when* and *where* (scene–time) and *what just happened* (sequence), aligning with its memory model.

4.3 Eight-Week In-Home Deployment: Longitudinal Adaptation and Experience (N=14)

4.3.1 Goals. We examined how SoundTrace fits within everyday lives over time, focusing on (1) *adaptation dynamics*—how recognition accuracy, user corrections, and scene dictionaries evolved through real use; (2) *trust formation and reliance*—how participants developed confidence and dependency on contextual feedback; and (3) *lived experience*—how SoundTrace intersected with accessibility, attention, and domestic or social routines.

We focused this field deployment on home environments, where environmental sounds play a central role in everyday awareness and safety for Deaf and hard-of-hearing individuals. Prior work has shown that the home is both acoustically rich and socially consequential for accessibility technologies [26, 27]. Domestic spaces afford stable contextual regularities (e.g., appliances, doors, alarms) that allow longitudinal observation of adaptation dynamics while limiting the confounds of uncontrolled public noise. Focusing on in-home use thus offered a tractable yet ecologically valid testbed for examining how contextual priors evolve, how users build trust over repeated interactions, and how SoundTrace fits into the rhythms of everyday domestic life.

4.3.2 Participants. Fourteen DHH participants (8 F / 6 M; ages 24–61, $M=40.9$, $SD=11.5$; Table 2) were recruited through social media, regional Deaf centers, and snowball sampling. Seven identified as hard-of-hearing, six as

Table 2. Field study participant demographics (N=14).

PID	Age	Gender	Identity	Hearing Loss	Onset Age	Hearing Device
P1	36	F	Hard-of-Hearing	Moderate	Birth	Bilateral Hearing Aids
P2	44	F	Deaf	Profound	Birth	None
P3	27	M	Hard-of-Hearing	Severe	Birth	Bilateral Hearing Aids
P4	50	F	Hard-of-Hearing	Severe	12 yrs	Single Hearing Aid
P5	29	F	Deaf	Profound	Birth	Cochlear Implants
P6	42	F	Hard-of-Hearing	Severe	8 yrs	Bilateral Hearing Aids
P7	31	F	Hard-of-Hearing	Moderate	10 yrs	Bilateral Hearing Aids
P8	24	F	Deaf	Profound	Birth	None
P9	61	M	Deaf	Profound	Birth	None
P10	33	M	deaf (small d)	Severe	Birth	Cochlear Implants
P11	47	M	Hard-of-Hearing	Moderate	15 yrs	Bilateral Hearing Aids
P12	39	M	Deaf	Profound	Birth	None
P13	56	F	Hard-of-Hearing	Moderate	Birth	Single Hearing Aid
P14	53	M	Deaf	Profound	Birth	Cochlear Implants

Deaf, and one as deaf. Five wore hearing aids in both ears, two wore a hearing aid in a single ear, three used cochlear implants, and four reported no assistive devices. Households included urban apartments (6), shared homes (4), and single-family houses (4). This sample was intentionally diverse but not statistically representative of the full DHH population. The study was IRB-approved; participants provided informed consent and received \$40/week as compensation.

4.3.3 Procedure and Instruments. Following a 30-minute onboarding (install, permissions, tutorial), participants used SoundTrace for eight weeks on Pixel-6 devices we provided. Participants were encouraged to use SoundTrace naturally and adjust foreground use, muting, and quiet hours according to household routines. Thus, the deployment examined user-controlled contextual awareness rather than assuming the system should remain salient at all times. We collected:

- **Usage logs:** daily foreground time, number of detections, per-day corrections, label volatility, scene-dictionary edits (rename/merge).
- **ESM prompts (3/week):** perceived usefulness of the most recent explanation (1–5), annoyance, and trust.
- **Weekly surveys:** trust, perceived accuracy, satisfaction (5-point Likert), open-ended reflections.
- **Interviews:** midpoint (Week 4, 30–45 min) and exit (Week 8, 60 min).
- **Diary snippets:** optional short notes about memorable interactions.

4.3.4 Analysis. We analyzed logs descriptively and longitudinally (per-participant and aggregate time series), and conducted *Applied Thematic Analysis* of interviews and free-text responses. Two coders independently open-coded five interviews to seed a codebook (usage rhythms, explanation practices, correction strategies, drift/recovery, attention management, social dynamics, safety/accessibility episodes, customization, and concerns). The codebook was iteratively refined; all interviews were then coded with dual pass. Inter-rater reliability across major themes achieved Cohen’s $\kappa=0.78$ and raw agreement=92.9%. We resolved disagreements through discussion, actively retained counterexamples, and triangulated themes with log and ESM data. We also constructed cross-case matrices to compare trajectories (early adopters vs. skeptics; single- vs. multi-occupant homes). As in the lab study, theme prevalence was summarized using both participant-level counts and coded-excerpt counts (Table 3).

Table 3. Theme prevalence in the eight-week field deployment (N=14).

Theme	Participants	Coded excerpts	% of excerpts
Shifting practices and growing reliance	13/14	74	25%
Learning routines and anticipating context	12/14	68	23%
Negotiating trust through explanations	11/14	49	16%
Social encounters and shared awareness	10/14	44	15%
Managing errors and ambiguity	8/14	38	13%
Privacy, placement, and perceived surveillance	6/14	28	9%

4.3.5 Findings. Our eight-week in-home deployment revealed how SoundTrace became deeply embedded in participants' everyday routines, shaped by evolving trust, contextual awareness, and negotiation. We organize our findings around six emergent themes that illustrate both the affordances and tensions of introducing a context-aware, memory-augmented recognition system into the intimate fabric of daily life.

Shifting Practices and Growing Reliance. Participants' use of SoundTrace evolved from curiosity to habitual reliance. In the first week, most described "testing" the system—playing sounds to probe recognition limits or verifying notifications manually. P3 recalled, *"At first I tried to confuse it—shut the door, bang pots, even run the blender. I wanted to see what it knew."* Over time, however, experimentation gave way to incorporation: by Week 4, 11 participants reported using SoundTrace passively, letting alerts run in the background while cooking, cleaning, or watching TV.

Usage logs reflected this shift. Mean daily active duration increased from 1.7 hours in Week 1 to 2.9 hours in Week 8. Participants described this not as "use" but "companionship"—P9 noted, *"I don't check it anymore; I just glance when it buzzes. It's like a quiet partner keeping watch."* This companionship extended to moments of reassurance: P5, who lived alone, said, *"Before, I used to worry if I missed something—a beep, a knock. Now I know I'll see it. That's relaxing."*

Interestingly, reliance varied by environment. SoundTrace was most valued in dynamic contexts (e.g., kitchens, shared living rooms), while participants in quieter homes sometimes paused notifications during rest hours. P11 explained, *"In the evening, I mute it; I don't need to know every creak."* This selective use suggests emerging literacy around balancing awareness and attentional load.

Participants emphasized that SoundTrace was most valuable for externally generated or otherwise missed sounds: knocks, door activity, appliance completion signals, alarms, and sounds produced by other people in shared spaces. Self-generated sounds, such as vacuuming or personal appliance use, were often treated as less important for immediate identification, although some participants still valued them for confirmation, logging, or retrospective review. This distinction suggests that context-aware recognition is most useful not when it labels everything, but when it helps users notice events that would otherwise remain ambiguous, unattended, or inaccessible.

Learning Routines and Anticipating Context. SoundTrace's contextual memory allowed it to develop temporal and spatial associations, which users came to recognize and anticipate. Several participants remarked that the system "remembered" their schedules. P8 described, *"By Week 3, it started showing 'microwave' every evening. I laughed—it learned my dinner time."* Such predictability strengthened user attachment, framing SoundTrace as "attentive" rather than mechanical.

Participants reported growing confidence not only in recognizing sounds but in understanding their timing. P6 reflected, *"I started expecting certain alerts... like laundry beep at night. If I didn't see it, I'd go check."* This reciprocal adaptation—users adjusting to system patterns, and vice versa—suggests an emergent rhythm of co-learning.

At times, contextual priors led to serendipitous discoveries. P2 noticed the system alerting to “door close” every morning at 7:45: *“It reminded me—oh, my neighbor leaves then. I never realized how regular it is.”* Such incidental insights reframed SoundTrace from assistive tool to environmental informant—making the invisible rhythms of domestic sound newly perceptible.

Yet this rhythm was not without drift. When participants moved between locations or changed network conditions, the system’s scene–time associations occasionally faltered. P10 recounted, *“It said ‘office’ but I was at home. Later I understood—it tagged by Wi-Fi.”* Similarly, P2 described a period of “confusion” after visiting a relative: *“At my sister’s place it was mixed up for a day. After I renamed it ‘New Home,’ it settled again.”* These moments of context drift, followed by recovery, revealed how users actively participated in retraining the system—renaming clusters, merging scenes, and resetting priors.

By Week 8, most participants had developed nuanced mental models of how episodic context shaped inference. Rather than viewing errors as failures, they interpreted them as signals of change—reminders that both human and system were adapting. This reflexive understanding illustrates a shift from using SoundTrace to collaborating with it: a distributed rhythm of sensing, forgetting, and relearning that mirrored the fluidity of home life itself.

Negotiating Trust Through Explanations. Trust in SoundTrace was not static; it was negotiated through cycles of verification, correction, and reflection. Explanations—particularly those citing location and time—served as key trust calibration tools. P4 shared, *“When it said ‘garage, morning, similar to car start,’ I thought yes—that fits. It shows it’s thinking.”* Such “seeing the reasoning” made errors more forgivable: P12 observed, *“It guessed ‘fan’ for heater, but said ‘evening, cold day’—I saw the logic. It wasn’t dumb.”*

Participants described “trust moments”—incidents that cemented confidence. P7 recounted a night when SoundTrace alerted to “glass clinking” from the kitchen: *“I went to look... My cat knocked a cup. It was right! After that, I trusted it.”* These episodes built emotional bonds; 10 participants personified the system (“it learns,” “it remembers”), aligning with prior findings on relational trust in situated AI [53].

However, over-trust occasionally emerged. P5 admitted, *“Sometimes I believe it too much—I saw ‘microwave’ but it was my phone.”* This blind trust underscores the double-edged role of contextual inference: while stabilizing perception, it may reduce vigilance. Two participants (P1, P13) requested optional uncertainty prompts (“Are you sure?”) to remind them of fallibility.

Social Encounters and Shared Awareness. Beyond individual interaction, SoundTrace shaped social dynamics. In multi-occupant homes, it became a shared awareness tool. P9’s family began interpreting alerts together: *“My kids shout, ‘It says doorbell!’ They join in.”* This communal use fostered inclusion; P9 said, *“Now I’m part of the moment, not catching up later.”*

Some participants used SoundTrace to mediate communication with hearing family. P3 noted, *“Before, I’d ask, ‘What was that?’ Now I show the app. It’s easier.”* Others leveraged logs as “proof” during disagreements—P8 recounted, *“My partner said the timer didn’t beep. I showed him the log!”*

Although SoundTrace was deployed in homes, several participants described how its benefits and challenges extend into semi-work or shared spaces. P2, who often performed manual tasks in a home workshop, noted: *“It caught the forklift beep I miss with earplugs.”* Yet, social visibility brought stigma concerns. P14 explained: *“People stare if it vibrates too much. I don’t want to explain.”* Balancing utility and social acceptability remains key for assistive tech adoption [28, 50].

Managing Errors and Ambiguity. Although accuracy improved over time (Section 4.1), participants encountered and negotiated errors. Misclassifications often prompted curiosity rather than frustration, especially when accompanied by plausible explanations. P6 observed, *“It said ‘knock’ but showed low confidence. I corrected it calmly.”*

Correction behavior revealed learning trajectories. Early in the study, participants corrected almost reflexively (*“I fix every wrong one,”* P11). Later, many adopted selective correction strategies, focusing on repeated or high-impact errors: *“If it keeps mixing fan and AC, I teach it. Minor ones I ignore.”* (P5)

Ambiguity was particularly salient for overlapping events. P10 described a complex kitchen scene: *“Blender, kettle, and water running—it just said ‘kitchen noise.’ I forgave it. Even I can’t separate them.”* Some appreciated this humility; P7 said, *“Better ‘unsure’ than wrong.”*

Importantly, SoundTrace’s explanations reframed errors as learnable: *“It’s like it’s thinking out loud. When it’s wrong, I see where it guessed.”* (P4) Participants viewed corrections not as burdens but as dialogue—*“like training a dog”* (P3)—underscoring the value of participatory adaptation.

Privacy, Placement, and Perceived Surveillance. While SoundTrace stored no raw audio, perceptions of privacy varied. Some participants valued the transparency of symbolic traces: *“I see what it remembers—just words, not sound. That’s fine.”* (P2) Others anthropomorphized memory, imagining “listening” even when muted. P11 confessed, *“I know it’s safe, but I still cover the mic sometimes.”*

Device placement emerged as both practical and symbolic. Participants experimented with how and where to keep their phones—on kitchen counters while cooking, propped near entryways when expecting deliveries, or carried in pockets during chores—to balance coverage, discretion, and convenience. P8 described this iterative adaptation: *“Week 1 I just left it on the couch—too far from the kitchen. By Week 3 I kept it by the hallway charger—perfect middle ground.”* Such adjustments reveal embodied negotiation of mobility, visibility, and audibility.

Some participants used SoundTrace beyond intended scope. P13 reported using it as a “sleep companion”: *“At night, if I hear vibration, I know something happened. It’s comforting.”* Others, however, preferred temporal boundaries: *“I turn it off at night. It’s my quiet time.”* (P12)

Participants also developed selective-use practices around repetitive audio such as television, music, and routine household noise. Rather than treating continuous use as universally desirable, participants adjusted quiet hours, muted low-value contexts, or relied on history review after the fact. These practices indicate that SoundTrace should be understood as a configurable awareness tool rather than an always-interrupting monitor.

Social privacy surfaced in shared spaces. P5 worried housemates might infer habits from the log: *“It shows when I cook or leave. That feels personal.”* Conversely, P9 found logs empowering: *“I show my care worker I didn’t miss the bell—they believe me.”* These tensions highlight the dual role of memory: as reassurance and as traceability.

Summary. Across eight weeks, SoundTrace transitioned from novel assistive app to embedded companion. Participants’ relationships with the system deepened through co-learning, trust calibration, and contextual awareness. Usage shifted from manual verification to ambient reliance; explanations transformed confusion into comprehension; and memory traces fostered both confidence and reflection. Yet these affordances coexisted with new negotiations—of social identity, privacy, and control. The findings underscore the complexity of designing *context-integrated assistive AI*: systems that must not only recognize sounds but inhabit the lived ecologies of their users.

5 Discussion

Across our controlled, lab, and field studies, SoundTrace demonstrated that contextual inference and lightweight memory jointly enhance both technical aspects and experiential qualities of sound awareness. Contextual priors reduced flicker and ambiguity, explanations and corrections fostered sensemaking and agency, and long-term use revealed how light-touch stewardship led to stable, personalized behavior. Building on these findings, we now situate SoundTrace within prior work and discuss how context, memory, and co-adaptation reframe assistive AI as a situated and participatory form of sensing.

5.1 From Context-Free to Context-Integrated Inference

Early work in sound awareness for deaf and hard-of-hearing (DHH) users has largely framed inference as an acoustic classification task, isolating events from their surroundings to achieve generalizable accuracy (e.g.,

SoundWatch [28], UbiEar [51]). While effective in controlled settings, such context-free models often struggle to maintain stability and relevance in the dynamic, noisy ecologies of daily life.

SoundTrace reframes inference as a contextual process, integrating scene, time, and sequence priors to ground predictions in ongoing activity. The mechanistic contribution of SoundTrace lies in separating acoustic evidence from contextual evidence while allowing both to jointly shape the posterior. The acoustic model captures immediate evidence (“what does this sound like?”), while scene–time and sequence priors encode longer- and shorter-term regularities (“what usually happens here now?” and “what tends to happen next?”). This separation allows contextual evidence to influence the posterior before temporal smoothing, distinguishing SoundTrace from post-hoc stabilization approaches and enabling both adaptability and interpretability. This design aligns with broader trends in memory-augmented systems in HCI [38, 54], which use episodic traces to scaffold reasoning. SoundTrace extends this idea from retrospective recall to real-time inference: it remembers in order to interpret the present.

Our quantitative results show that context integration improves both accuracy and volatility, particularly under noise, aligning with work in machine learning that frames context as a form of regularization rather than bias [8, 32]. Beyond performance, participants experienced this stability as a social and cognitive property: predictions that “made sense for the moment” inspired trust and reduced cognitive load. This resonates with prior work on situated sensing and context-aware systems (e.g., [1]), which emphasizes coherence with environmental context and expectations of place. In this light, SoundTrace shifts the design space from sound recognition toward *sound reasoning*—from labeling isolated events to interpreting them within a lived temporal frame.

5.2 Memory as Situated Adaptation

Memory has long figured in assistive technology research, from retrospective archives for reflection (e.g., Memoro [54]) to adaptive personalization systems that adjust over time [9, 16]. SoundTrace contributes a complementary notion: *episodic continuity*. Its lightweight symbolic traces—scene-time event tuples—serve not as records to recall later but as scaffolds for ongoing inference and explanation. In doing so, the system embodies what Suchman [52] described as *situated action*, adapting to local contingencies rather than following static rules.

Participants’ descriptions of SoundTrace “remembering routines” or “settling into my kitchen sounds” reveal that memory, when made visible, becomes a site of collaboration. Users not only experienced adaptation but actively participated in it, renaming clusters and correcting events to re-align the system’s sense of place and time. This reflects a broader shift in HCI toward *mutual adaptation* rather than unidirectional learning [42, 43]. As in other contexts of everyday AI (e.g., smart thermostats, routine-predictive assistants), the value of memory lies not in precision but in continuity—the sense that the system evolves with the rhythms of daily life.

This finding extends theories of *intimate infrastructure* [6, 37], positioning memory as a material through which users domesticate AI. Rather than static personalization parameters, episodic memory becomes a shared interpretive resource: one that can be inspected, renamed, or reset. SoundTrace thus invites us to reconceptualize memory in assistive systems not merely as a storage function but as a *temporal boundary resource*—a means for people and models to stay in sync as both lives and contexts change.

5.3 Explainability as Co-Adaptation

Explainable AI has been widely studied in domains where users audit, debug, or contest algorithmic outcomes [31, 35, 44]. In assistive technologies, however, explanations serve a different role: they help users *align perception* rather than evaluate fairness. SoundTrace’s natural-language rationales (“door, evening, after footsteps”) exemplify what Lim et al. [35] called *everyday intelligibility*—the kind of transparency that supports understanding-in-action.

Our lab and field studies showed that explanations in SoundTrace functioned as *trust calibration* devices. Participants used them to judge plausibility (“It fits—garage, morning, car start”) and to interpret misclassifications

as understandable rather than arbitrary. This aligns with Kulesza et al.’s work on *explanatory debugging* [31], yet extends it beyond isolated tasks into continuous, embodied interaction. Here, explanation and correction together formed a *co-adaptation loop*: as users corrected labels, the system stabilized; as the system stabilized, users corrected less, reflecting mutual learning.

Crucially, SoundTrace shows that explanations need not always increase detail; often, brevity and contextual relevance matter more. Participants preferred one-sentence rationales tied to time or sequence cues, echoing prior findings on *intelligibility-on-demand* [34]. Such lightweight explainability is vital in mobile assistive contexts, where attention and trust are dynamically negotiated. In this sense, explanation in SoundTrace operates less as a window into the model and more as a *conversational cue*—maintaining joint attention between user and system in the flow of daily activity.

5.4 Negotiating Privacy, Agency, and Visibility

While SoundTrace’s memory architecture mitigated many privacy concerns by avoiding raw audio retention, participants’ reactions underscored that privacy is *experienced*, not merely *ensured*. Symbolic traces were perceived as safe by some and intrusive by others, depending on household composition and social norms. The phone’s placement—on kitchen counters, in pockets, near doorways—became an ongoing negotiation between convenience, coverage, and discretion.

These embodied choices reveal that visibility and privacy in assistive sensing are *co-constructed through everyday practice*, not dictated by technical safeguards alone. As prior work on situated privacy and domestic IoT has shown [2, 37], people manage exposure not through binary permissions but through *micro-placements* and *temporal boundaries*. Participants’ gestures—covering the mic, renaming scenes, disabling alerts during rest hours—illustrate privacy as a *lived calibration of agency*.

Future assistive systems could make this negotiation more explicit through *selective memory views* or *ephemeral logs*, allowing users to see, prune, and forget traces over time. Designing for *forgetfulness*, as much as for recall, might better align with the rhythms of domestic life, where reassurance and invisibility must coexist.

5.5 Designing for Salience, Change, and Everyday Ecology

Across longitudinal use, participants developed a literacy for negotiating which sounds mattered—a shift from total awareness toward *selective salience*. Some muted repetitive events; others corrected only high-impact sounds. This emergent prioritization highlights an open frontier in assistive sound recognition: beyond “what is happening,” users need to know “what is new, missing, or different.” Although SoundTrace did not explicitly model difference or novelty, participants themselves enacted these capacities through sensemaking. By interpreting stable patterns and noticing when expected cues were absent or anomalous (“it didn’t show the microwave today”), they effectively performed *difference detection* in collaboration with the system. In this sense, SoundTrace already instantiated change-awareness conceptually through human–AI co-adaptation, even if not yet algorithmically realized. Future systems might build on this foundation by modeling not just sound categories but *temporal deviation*—detecting changes against learned baselines while preserving the participatory, situated character of interaction.

Such an approach resonates with broader possibilities in contextual AI. For blind and low-vision users, scene-recognition technologies can emphasize difference detection—for instance, identifying what has changed in a workspace or who has entered a room [20, 36]. For neurodivergent users or those with cognitive impairments, lightweight environmental memory could surface routine disruptions (“the oven is still on,” “the door was left open”) as gentle prompts rather than alarms. Beyond accessibility, domestic sensing technologies might similarly leverage episodic continuity to support wellbeing (“the house is quieter than usual”) or safety (“unusual motion near the door”).

In each case, SoundTrace’s approach—context-integrated, interpretable, and user-correctable—offers a foundation for designing *participatory sensing ecologies*: systems that learn what users consider meaningful through ongoing dialogue rather than predefined salience models. Designing for *change awareness* rather than omniscience could make assistive AI less interruptive, more respectful, and ultimately more attuned to the lived temporalities of home and work.

5.6 Intended Use and Failure Modes

SoundTrace is designed for user-controlled contextual sound awareness rather than mandatory always-on monitoring. In practice, participants used the system in multiple modes: ambient use during active household routines, on-demand checking when they wanted to verify or review a sound, and quiet-hours use when they wanted reduced interruption. This flexible model is important because the value of recognition depends on activity, setting, and attention. A kitchen timer, entryway knock, or appliance completion signal may be highly valuable when the user is elsewhere, while a self-generated sound may be less useful for immediate identification.

SoundTrace is most effective when contextual regularities are moderately stable and when scene labels are meaningful to the user. It can fail or provide limited value when routines change abruptly, when the phone is far from the sound source, when overlapping events occur at similar intensity, when Wi-Fi/GPS context is unavailable or misleading, or when contextual priors reinforce a previously learned error. It may also be less useful in highly repetitive acoustic environments, such as prolonged television audio, unless users configure filtering or quiet hours. To mitigate these issues, SoundTrace uses confidence thresholds, exponential decay, correction-driven updates, scene editing, quiet hours, and memory reset. These mechanisms do not eliminate failure, but they give users tools to recover from drift and align the system with changing routines.

5.7 Limitations and Future Work

As with all empirical work, our studies were conducted under bounded and controlled conditions that constrain generalizability. The field deployment focused on domestic settings, where routines and acoustic regularities are strongest; how contextual priors behave in more fluid or unpredictable settings—such as workplaces, transit, or outdoor spaces, which frequently present overlapping sounds or less stable routines—remains an open question. Future research should examine mobility, overlapping social contexts, and multi-device or distributed sensing configurations that more fully represent everyday diversity.

The DHH participants in our studies do not represent the full diversity of DHH communities; future deployments should include larger samples, signing-first users, older adults, children and families, and use settings beyond the home. Similarly, while eight weeks allowed us to observe adaptation and trust formation, longer-term studies are needed to understand model drift, user fatigue, and the persistence of personalized routines beyond the study window.

Although SoundTrace achieved robust on-device performance, it did not explicitly model *sound salience* or *event priority*. Addressing this limitation could enable systems that adapt not only to context but to user goals and attentional states. Similarly, our memory representation was symbolic and local; richer multimodal traces (e.g., low-rate embeddings or summaries of concurrent visual context) could enhance continuity while preserving privacy.

Finally, participants often described SoundTrace as a “companion,” a framing that invites both opportunity and caution. The relational dynamics of assistive AI—trust, dependence, and emotional attachment—merit deeper longitudinal study. Future work should examine when companionship becomes over-reliance and how uncertainty cues, decay mechanisms, or periodic recalibration might sustain healthy human–AI relationships over time.

6 Conclusion

SoundTrace advances assistive sound awareness by demonstrating that memory-augmented, context-integrated inference can function as a real-time, interpretable, and user-correctable system for everyday environments—transforming sound recognition from a momentary act of classification into an ongoing, situated process of sensemaking. Through controlled, lab, and in-home studies, we showed that embedding temporal and spatial priors improves robustness and coherence, while exposing symbolic memory and explanations fosters agency, personalization, and trust. Beyond technical performance, these findings advance a broader perspective on assistive AI as a form of participatory sensing—one that learns with users through continuity and correction rather than optimizing in isolation. Future work can extend this framework toward richer multimodal traces and adaptive salience modeling, exploring how human–AI co-adaptation might support awareness, reflection, and care across increasingly diverse everyday ecologies.

Generative AI Disclosure. Generative AI was used to create images for Figure 1 and for limited text editing to improve grammar and clarity. All scientific content, analyses, and conclusions are the authors' own.

References

- [1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Handheld and Ubiquitous Computing*, Hans-W. Gellersen (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 304–307.
- [2] Wael Albayaydh and Ivan Flechais. 2024. "Innovative Technologies or Invasive Technologies?": Exploring Design Challenges of Privacy Protection With Smart Home in Jordan. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–54.
- [3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.
- [4] Sara Barahona, Diego de Benito-Gorrón, Doroteo T. Toledano, and Daniel Ramos. 2024. Enhancing Conformer-Based Sound Event Detection Using Frequency Dynamic Convolutions and BEATs Audio Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 3896–3907. doi:10.1109/TASLP.2024.3444490
- [5] 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* (Reno Nevada USA, 2016-10-23). ACM, 3–13. doi:10.1145/2982142.2982171
- [6] Joan Cahill, Raul Portales, Sean McLoughin, Nithia Nagan, Braden Henrichs, and Sean Wetherall. 2019. IoT/sensor-based infrastructures promoting a sense of home, independent living, comfort and wellness. *Sensors* 19, 3 (2019), 485.
- [7] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. BEATs: Audio Pre-Training with Acoustic Tokenizers. arXiv:2212.09058 [eess.AS] <https://arxiv.org/abs/2212.09058>
- [8] Gabriele Civitarese, Riccardo Presotto, and Claudio Bettini. 2019. Context-driven active and incremental activity recognition. *arXiv preprint arXiv:1906.03033* (2019).
- [9] Caitlyn Clabaugh, Kartik Mahajan, Shomik Jain, Roxanna Pakkar, David Becerra, Zhonghao Shi, Eric Deng, Rhianna Lee, Gisele Ragusa, and Maja Mataric. 2019. Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Frontiers in Robotics and AI* 6 (2019), 110.
- [10] Hang Do, Quan Dang, Jeremy Zhengqi Huang, and Dhruv Jain. 2023. AdaptiveSound: An Interactive Feedback-Loop System to Improve Sound Recognition for Deaf and Hard of Hearing Users. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2023-10-22) (ASSETS '23). Association for Computing Machinery, 1–12. doi:10.1145/3597638.3608390
- [11] Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. 2025. Mirai: A Wearable Proactive AI" Inner-Voice" for Contextual Nudging. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.
- [12] Cathy Mengying Fang, Wazeer Zulfikar, Yasith Samaradivakara, Suranga Nanayakkara, and Pattie Maes. 2025. The Goldilocks Time Window for Proactive Interventions in Wearable AI Systems. *arXiv preprint arXiv:2504.09332* (2025).
- [13] Jort F Gemmeke, Daniel PW 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.
- [14] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of Interspeech*. 571–575. doi:10.21437/Interspeech.2021-698

- [15] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. 2022. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10699–10709.
- [16] Suat Gonul, Tuncay Namli, Sasja Huisman, Gokce Banu Laleci Erturkmen, Ismail Hakki Toroslu, and Ahmet Cosar. 2018. An expandable approach for design and personalization of digital, just-in-time adaptive interventions. *Journal of the American Medical Informatics Association* 26, 3 (dec 2018), 198–210. doi:10.1093/jamia/ocy160
- [17] 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 (2021), 1–23.
- [18] 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 *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [19] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied Thematic Analysis*. Sage Publications.
- [20] Enqiang Guo, Xinsha Fu, Jiawei Zhu, Min Deng, Yu Liu, Qing Zhu, and Haifeng Li. 2018. Learning to measure change: Fully convolutional siamese metric networks for scene change detection. *arXiv preprint arXiv:1810.09111* (2018).
- [21] F. Wai-ling Ho-Ching, Jennifer Mankoff, and James A. Landay. 2003. Can you see what i hear? the design and evaluation of a peripheral sound display for the deaf. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) (CHI '03). Association for Computing Machinery, New York, NY, USA, 161–168. doi:10.1145/642611.642641
- [22] 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 *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2023-10-22) (ASSETS '23). Association for Computing Machinery, 1–14. doi:10.1145/3597638.3608431
- [23] Jeremy Zhengqi Huang, Jaylin Herskovitz, Liang-Yuan Wu, Cecily Morrison, and Dhruv Jain. 2025. Weaving Sound Information to Support Real-Time Sensemaking of Auditory Environments: Co-Designing with a DHH User.
- [24] Jeremy Zhengqi Huang, Reyna Wood, Hriday Chhabria, and Dhruv Jain. 2024. A human-AI collaborative approach for designing sound awareness systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [25] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022-04-29) (CHI '22). Association for Computing Machinery, 1–16. doi:10.1145/3491102.3502020
- [26] 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*. 1–13.
- [27] 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* (New York, NY, USA, 2020-04-23) (CHI '20). Association for Computing Machinery, 1–12. doi:10.1145/3313831.3376758
- [28] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. SoundWatch: Exploring Smartwatch-based Deep Learning Approaches to Support Sound Awareness for Deaf and Hard of Hearing Users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event Greece, 2020-10-26). ACM, 1–13. doi:10.1145/3373625.3416991
- [29] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Khoa Nguyen, Rachel Grossman-Kahn, Leah Findlater, and Jon Froehlich. 2022. SoundWatch: Deep learning for sound accessibility on smartwatches. *Commun. ACM* 65, 6 (2022), 100–108.
- [30] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 28 (Nov. 2020), 2880–2894. doi:10.1109/TASLP.2020.3030497
- [31] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- [32] Benjamin Lengerich, Caleb N Ellington, Andrea Rubbi, Manolis Kellis, and Eric P Xing. 2023. Contextualized machine learning. *arXiv preprint arXiv:2310.11340* (2023).
- [33] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).
- [34] Brian Y Lim and Anind K Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*. 415–424.
- [35] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.
- [36] Chun-Jung Lin, Sourav Garg, Tat-Jun Chin, and Feras Dayoub. 2025. Robust scene change detection using visual foundation models and cross-attention mechanisms. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8337–8343.

- [37] Heather Richter Lipford, Madiha Tabassum, Paritosh Bahirat, Yaxing Yao, and Bart P. Knijnenburg. 2022. Privacy and the Internet of Things. In *Modern Socio-Technical Perspectives on Privacy*, Bart P. Knijnenburg, Xinru Page, Pamela Wisniewski, Heather Richter Lipford, Nicholas Proferes, and Jennifer Romano (Eds.). Springer International Publishing, Cham, 233–264. doi:10.1007/978-3-030-82786-1_11
- [38] Natasha Maniar, Samantha WT Chan, Wazeer Zulfikar, Scott Ren, Christine Xu, and Pattie Maes. 2025. MemPal: Leveraging multimodal AI and LLMs for voice-activated object retrieval in homes of older adults. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 993–1015.
- [39] 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 (2006), 333–351.
- [40] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* 102, 3 (1995), 419.
- [41] Elizabeth D Mynatt, Maribeth Back, Roy Want, and Ron Frederick. 1997. Audio Aura: Light-weight audio augmented reality. In *Proceedings of the 10th annual ACM symposium on User interface software and technology*. 211–212.
- [42] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. 2016. Formalizing Human-Robot Mutual Adaptation: A Bounded Memory Mode. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 75–82.
- [43] Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. 2017. Human-Robot Mutual Adaptation in Shared Autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 294–302.
- [44] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of explanation conceptualisations on trust in AI-assisted credibility assessment. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–31.
- [45] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.
- [46] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1015–1018.
- [47] Lawrence R Rabiner. 2002. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (2002), 257–286.
- [48] Justin Salamon, Christopher Jacoby, and Juan P. Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1041–1044.
- [49] Nitin Sawhney and Chris Schmandt. 1999. Nomadic radio: Scaleable and contextual notification for wearable audio messaging. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 96–103.
- [50] Kristen Shinohara and Jacob O Wobbrock. 2011. In the shadow of misperception: assistive technology use and social interactions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 705–714.
- [51] Liu Sicong, Zhou Zimu, Du Junzhao, Shangguan Longfei, Jun Han, and Xin Wang. 2017. UbiEar: Bringing location-independent sound awareness to the hard-of-hearing people with smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–21.
- [52] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
- [53] Rongbin Yang and Santoso Wibowo. 2022. User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets* 32, 4 (2022), 2053–2077.
- [54] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

A Additional Results

Table 4. Performance comparison across different noise levels (SNR in dB), measured using Macro-F1 (%).

Model	Clean	0 dB	-5 dB	-10 dB
Baseline (AST-lite)	71.4	63.2	56.7	48.8
+Scene	75.3	69.5	65.1	59.0
+Seq	74.6	68.7	64.4	58.1
Full (SoundTrace)	78.1	74.9	70.8	66.3

Table 5. Backbone comparison showing consistent gains from contextual inference.

Backbone	Context Fusion	Macro-F1	Volatility
CNN	No	68.2	0.36
CNN	Yes	73.0	0.27
AST-lite	No	71.4	0.32
AST-lite	Yes	78.1	0.22
BEATs-lite	No	74.8	0.29
BEATs-lite	Yes	79.6	0.21