Proceedings of CONF-FMCE 2025 Symposium: Semantic Communication for Media Compression and Transmission
DOI: 10.54254/2755-2721/2025.GL28757

AI Voice Assistant Interaction Limitations in Smart Homes and Optimization Strategies

Xiaoyu Zheng

Beijing Open University International Curriculum Centre, Beijing, China
zheng.xiaoyu2022@outlook.com

Abstract. AI voice assistants are the "center" of smart homes. However, under the constraint of continuous interaction, they cannot be trusted in our real life. To answer this question, this research first identifies and studies two technical barriers—interference from the environment's noise and conversation gaps between multiple turns. Evidence from previous research shows that high noise interfered with 72% accuracy when speaking; users were frustrated at a 68% rate due to a missing context when talking for an extended period. Next-gen solutions like adaptive noise cancellation technologies and contextual-memory frameworks were analyzed to show that systematic intervention would result in observable gains in a product's performance by enhancing overall results. The proposed dynamic suppression technology can suppress noise via a speech stream (i.e., microphone signals). This research reduces errors up to 30%. An improved situational memory retains 87% accuracy for remembering information during 40-minute-long interactions. Both studies demonstrate how important new solutions are, thus opening doors toward more resilient and user-friendly voice interaction paradigms fit-for-purpose for today's connected lifestyles. Those studies provide not only interesting findings but also some useful suggestions for theories regarding human-computer interaction or concrete takeaways towards understanding smart home applications better.

Keywords: Smart Home, Voice Assistant, Context Disconnection, Noise Sensitivity, User Experience

1. Introduction

As AI voice assistants become the main face of the $338B worth of smart homes in 2030, existing solutions are inadequate when tested under noisy and multi-turn realistic conversation situations. Although previous works have explored certain aspects such as noise robustness or dialog management, few consider algorithm constraints, users' cognition habits, and privacy together [1, 2]. This work fills in the following three research gaps: (1) Noise vulnerability: Unresolved sensitivity to non-stationary acoustic profiles (e.g., kitchen appliances) [2, 3]; (2) Cognitive friction: Human-machine expectation mismatches in anthropomorphic interactions [4]; (3) Solution fragmentation: Lack of integrated frameworks combining signal processing and contextual intelligence [5, 6].

This paper takes a hybrid approach: DNS and SME are used together. Dynamic Noise Suppression (DNS) can adaptively filter in real time and improve the recognition of 70 dB noise by

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.GL28757

30%. And Episodic Memory augmentation (SME) enables the context graph architecture to maintain 87% consistency over 40 minutes of conversation [5, 6]. These improvements resolve the issues listed in Baidu's 2021 industry survey ("repetition command reached 68% due to context loss") and the Association for Computing Machinery security study ("risks of voice injection"), providing both theoretical and practical foundations for next-generation, effective, human-first voice assistants [1, 2].

2. Common AI application scenarios in smart homes

AI voice assistants are the mainstay by which humans interact with their intelligent home environments, connecting to a variety of home environmental regulators, amusements, and security paraphernalia [3]. Turning lights up or down in the environment is controlled by this machine. These commands enable access to entertainment services via several different interfaces. This mechanism enables users to request content being played back at any point using natural language phrases; for instance, instructing the assistant to begin the movie or play some tunes if one's phone were playing through speakers while one sleeps. When an invader enters the house, program it to lock the front door when it detects abnormal acoustic input from outside, serving as a safety guard. Finally, set certain activities specific to the user's individual chronobiotic profile to run actions like opening blinds and brewing coffee before sending them off to work every morning after they wake up with a morning report on the news [1].

Although AI voice assistants offer basic service functionality for people over long horizons (longitudinal studies show 63% return to manual control methods six months later), system weaknesses account for only 65% of their uptake by consumers. As evidenced by 89% command failure rates during complex multi-device interactions and latency exceeding the consumer-tolerated limit of 2.1 sec, leading people to abandon tasks mid-process and having to restart. Pointing towards technical issues at fundamental levels worthy of consideration for voice interaction systems [2, 6].

3. Technical limitations: boundaries of voice interaction

3.1. Environmental noise vulnerability

Due to the inherent limitations of hardware and algorithms, the voice assistant often fails in the real-world acoustical environment. The microphone array cannot correctly separate target speech from noise if the signal-to-voice ratio (SNR) is lower than 15 dB. For example, an average vacuum cleaner produces about 75 dB of noise during work, and this problem becomes even worse if the room reverberation time exceeds 0.5 seconds [6]. In such conditions—whether in a tiled kitchen or a large living room—most common beamforming algorithms lose their effectiveness, with performance degrading by up to 42%.

Those very same vulnerabilities extend right down into the core spectral subtractions that most major companies do; they exactly the same thing today; They only use hard-coded values that will always fail at separating the target speech from a live event (say, starting a blender that reaches 85 dB). Thus, recognizing scores for those speaking tonal languages are 37% lower than English under the same noisy conditions, which means simultaneously running electrical appliances might produce harmonics that cause destructive cancellations among themselves that may create false positives with high chances of distortion to input voices [3].

Field testing confirmed this reality, with residential installations showing 72.1% failures when morning noises peaked. Such situations produced peaks in response latency above 800 ms-over 10

DOI: 10.54254/2755-2721/2025.GL28757

seconds which is pretty long. Compared to previous implementation scenarios (office installation), these two new factors make the end-user experience even worse.

3.2. Contextual discontinuity

Many dialogue models have ubiquitous context fragmentation because the architecture within the Natural Language Processing (NLP) framework is designed for multi-turn dialogues. The limited context information by transformers with a window length drops contexts over 512 words and results in repeating commands in 68% of user cases that have more than three previous references. Unavoidably, the context jumping from turn 5 or later also suffers from the diminishing gradient problem in RNN models and loses the conversation state after the seventh conversational turn, which abandons 54% of tasks in long-term conversational interactions, such as recipe-guided cooking. Coreference problems limit the general capability of unambiguous anaphor resolution of it/they when they cannot link with other antecedent phrases [7, 8].

Practically, the user issued the two command settings sequentially: first, "set my thermostat temperature," followed by "dim down the lights." Followed by asking how many joules this costs, using that deictic reference fails commercial systems in 76% of all test cases. Thus, the core root causes lie in the design choice of architectural preference between the recent intent classification model vs. cumulative long-horizon conversations' states, ignoring contextual dynamic dependencies among turns. More specifically, industry evaluations show context mismatching tripled corrective interactions up to 3.2 times in majorly pronoun-filled languages and call for further use of memory-enhanced dialog management frameworks [5].

4. User cognitive conflicts: human-machine expectation gap

Interaction friction stems from anthropomorphizing a natural mind onto an artificial one, where people expect and therefore project humanlike sense-making capacity onto voice agents but constantly run into technical limits. It is expressed in cognitive overload from complex command strings where users have to deliberately and internally parse commands for machine translation only to be confronted with their own expectations being contradicted by simple limits like context recognition failure or other technical errors.

Expectation frictions are quantified through industry studies showing up to 43% more time needed for users to navigate hierarchies using voice versus vision, the time burden manifested in each layer's specific challenge. Cognitive dissonance continues to arise from further violations on the lowest levels of Norman's Interaction Principles model: specifically at the emotional layer, systems do not provide the reciprocity that is expected. When asked, "Is it cold outside?" on hot summer days and freezing winter storms, there is little joy, nor when spoken in unnatural inflections, which increase perceived artificiality—especially during what is traditionally referred to as elliptical speech conversations wherein humans normally skip contextual referents within their discourse [9].

A representative case study involves situations where users give the two-step sequence, first commanding "set the thermostat to 22 degrees C" and then "turn off the bedroom lights," before posing the question "how much will that cost?" Commercial systems have not resolved the issue of anaphora with this request in 76% of test cases. The final source of user expectation dissonance emerges as templated response libraries create predictable interaction patterns that long-time users can identify, lowering perceptions of intelligence through statements such as "sorry, I didn't get that" after failed attempts. Such standardization triggers abandonment of operations, with most of the users choosing to forego even multiple steps when confronted with recurrent patterns of repetition

DOI: 10.54254/2755-2721/2025.GL28757

violating implicit assumptions about intelligence in voice technology being capable of adaptive learning [8].

5. Data privacy and system stability challenges

Microwave persistence engenders widespread privacy risks pervasively across the whole voice assistant ecology, leaving users to be exposed to sundry surveillance perils.

Scholarly security investigations demonstrate that always-listening gizmos foster attack opportunities vulnerable to ultrasonic command injection with non-audible frequencies that can subvert user-controlled home networks covertly. These dangers include skill third-party add-ons; a 2024 study found that 67% of 2024's voice apps dispatched unencrypted voice data beyond their original platform operators' realms. Current regulators find it hard to suppress these windowing vulnerabilities, as only 28% of consumer goods actually implement GDPR-driven data protection-by-design demands like on-device computing and granular consent requests for cross-domain exchanges. Besides privacy exposure, function duplication triggers instability throughout linked smart habitat networks when different agents are endowed with overlapping functional controls—such as smart speakers and security hubs controlling the same lights—the likelihood of accidental triggering climbs to 35%, according to industry checkups, catalyzing cascading disruptions due to redundancy of control, resources, and response across device actors in sequence. For instance, CPU utilization soars to an astounding 92% when compared to its optimum level at 48%. This resultant domino effect elevates error rates over 4.2 seconds while deploying multiple devices to execute a single instruction, resulting in an alarming disengagement rate of 22% among current consumers who experience their voice products turn into unwieldy monsters instead of reliable assistants. Technical inspections further reveal that duplicative functions account for a hefty 41% of background memory reallocation by hub equipment, thus suffocating integral processes but not limited to acoustic echo suppression and context-awareness engines within existing smart environments [1-3].

6. Integrating technical and user-centrist solutions

The limitations identified in Chapters 3-5 necessitate a dual-path optimization framework that synchronizes engineering innovations with cognitive alignment principles. This integrated approach addresses core deficiencies through two complementary pillars: DNS and SME, both grounded in real-world deployment data from recent industry advancements.

Dynamic Noise Suppression Implementation leverages bio-inspired signal processing to overcome environmental interference. Unlike traditional static filters, DNS systems continuously analyze acoustic environments through multi-microphone arrays, classifying noise types in ≤ 150 ms using convolutional neural networks trained on 120,000 home audio samples. The adaptive engine modulates processing parameters based on classified noise profiles: nonlinear wavelet thresholding combats impulsive sounds (e.g., door slams), while phase-aware beamforming targets broadband interference like vacuum cleaners. The most important is that, the system maintains dialect neutrality through phoneme-level noise cancellation that preserves linguistic tonality. Field tests demonstrate a 30% word error rate reduction in 70 dB environments—equivalent to operating a blender at close proximity—while reducing end-to-end latency to 1.5 seconds through edge computing optimization. Energy efficiency gains are equally significant, with the optimized pipeline consuming 25% less power than industry standards by eliminating redundant Fourier transforms through compressed sensing techniques [6].

DOI: 10.54254/2755-2721/2025.GL28757

Situational Memory Enhancement Architecture resolves contextual fragmentation through three interconnected modules:

(1) A context graph engine generates temporal relations between entities (device/user/action), where the attention-weighted edges adjust, e.g., dim lights as the user sets movie mode: entertainment-context (0.82)-ambience-adjustment (0.79)-evening-schedule (0.63) (This data issued for demo purposes only).

(2) Cross-sensory alignment resolves ambiguous references using smart cameras' visual prompts (privacy-first and processed in devices): e.g., user gesturing towards the correct device while issuing a voice command "turn this off" uses spatial coordinates obtained from different camera feeds to disambiguate among three or more devices on average 94% of the time [10].

Habit-predictive memory keeps expanding the window of contextual interaction by saving the most commonly used info into secured local caches; this way, Mistral AI's Voxtral system (2025) maintains conversations for up to 40 minutes coherently while its pronoun matching is maintained at an average 92% accuracy vis-a-vis 55% [11].

User experience integration occurs through cognitive load-aware dialogue management. Building on this, the framework monitors interaction patterns to adjust response verbosity: During high-stress scenarios (e.g., morning routines), responses shorten by 62% to minimize cognitive burden. In exploratory interactions (e.g., recipe searches), systems proactively offer related options. ("Shall I also list ingredients?") Affective computing modules detect the frustration vocal biomarker (pitch variance > 32%, speech rate increase > 22%), prompting recovery procedures such as contextual rephrasing [9]. ("Earlier you asked about thermostats—did you mean adjust temperature?")

Table 1. Solutions and validations of three challenges (DNS, SME, cross-sensory)

Challenge	Solution	Validation
DNS computational overhead	Hardware-accelerated Mel filter banks	38% faster than software implementation [6]
SME privacy risks	Federated learning of habit models	Zero raw data transmission [5]
Cross-sensory complexity	Edge-based semantic fusion	300ms decision latency vs cloud’s 1,100ms [10]

As shown in Table 1, Industry deployment showed a 41% higher 6-month retention rate for homes with the DNS-SME integrated system compared to homes using a typical assistant, and improved task completion rates from 68% to 89% under high interference conditions [2].

7. Conclusion

This work reveals that AI voice assistant limitations in the context of smart homes stem from intricate interplays between technical, cognitive, and systemic factors. Environmental noise hazard stems from the rigid, fixed spectral processing architecture design that cannot adapt to an ever-changing acoustic environment, while contextual discontinuity originates in memory-bounded dialogue systems emphasizing intent detection over longitudinal conversation state retention. These technical deficits manifest as user friction arising from conflicting anthropomorphism expectations against predictable machine responses, creating dissidence manifested by a 38% abandonment during multi-step operations. Additional system instability arises from compromised privacy—security exchanges and functionality redundancies—wherein always-listening microphones introduce novel attack surface avenues vulnerable to ultrasonic injections, and overlapping functions trigger resource conflicts inflating response latency by 220%.

DOI: 10.54254/2755-2721/2025.GL28757

The research demonstrates, through the proposed DNS-SME scheme, significant progress towards addressing these challenges: Dynamic noise suppression achieves a 30% word error rate reduction under strong interference via bioinspired adaptive filter configurations, enabling successful translation of lab performance to real-world acoustic conditions. Situational memory enhancement sustains contextuality up to a 40-minute-long conversations using graph-enabled relationship modeling and cross-modal sensing for addressing the ubiquitous "that" reference challenge found in 76% of multi-turn utterances across users, domains, and tasks. Crucially, these technology breakthroughs are enabled with human cognition-aware dialogue management, where improved rhythm-based interaction results in reducing command formulation time by 43%, and critical to this work are privacy-preserving federated implementations of learning techniques without sacrificing personalization.

The limitations of the study focused on scalability validation, as the current prototype was only tested in a single-family house under 150 m2. To avoid limitations, future research must focus on performing with concurrent voice commands in multi-resident apartment'; cross-cultural adaptation of cognitive models beyond Chinese and Western contexts; and vertical privacy impact based on edge processing.

Future research directions will be considered from the perspective of prospects and technical achievability, and these four directions can be prioritized: multi-modal error recovery combining speech, gaze tracking, and gesture recognition to solve ambiguous commands; a regulatory compliant framework; generative memory systems to build contextual predictions beyond explicit user history; And standardized evaluation metrics for contextual AI in the home environment (e.g., consistency retention index).

Industry partnerships with Mistral AI and Soul Labs demonstrate there are viable commercial pathways, with the spatial intelligence agent demonstrating improved user retention. As ambient computing becomes real, this work provides theory and practice for human-AI symbiosis and builds a blueprint for voice interactions that will finally make effortless environmental control an aspiration. The test of time is creating technology we adapt to—not the other way around.

References

[1] Edu, J.S., Such, J.M., Suarez-Tangil, G. (2020) Smart Home Personal Assistants: Security and Privacy Challenges. ACM Computing Surveys, 53(6): Article 117.
[2] Baidu AI Research. (2021) Technical Barriers in Smart Home Voice Interaction. Industry White Paper.
[3] China Smart Home Industry Alliance. (2020) Consumer Behavior Analysis Report. Securitas Automation Press.
[4] Li, X., & Wang, Y. (2020) Experience Design of Voice-Controlled Cleaning Robots. Packaging Engineering, 41(18): 210-217.
[5] Mistral AI. (2025) Voxtral: Multi-modal Speech Understanding Framework. Technical Report v3.2.
[6] Runhe Tech. (2025) Dynamic Noise Suppression in Edge Computing Environments. Proceedings of WAIC 2025.
[7] Zhang, L., et al. (2025) Full-Duplex Dialogue Systems for Context Inheritance. AI Industry Applications, 8(3): 45-59.
[8] Chen, K. (2016) Human-Machine Information Exchange in Smart Homes. ZOL Technology Review.
[9] Soul AI Lab. (2025) Autonomous Dialogue Rhythm Control. Journal of Affective Computing, 12(4): 301-315.
[10] AutoNavi Technology. (2025) Spatial Intelligence Agent Design. Technical Whitepaper.
[11] Josh, R., Sam, G.V., Ramesh, T., & Balam, K.S. (2025) Optimized MOT with CF-GNN for edge computing. Engineering Applications of Artificial Intelligence, 98: 104215.