An Agentic Voice-Based Assistant for Interactive Conversation and Guidance in Real-World Environments

Demonstration Track

Donghao HUANG*
Mastercard
Arlington, VA, United States
donghao.huang@mastercard.com

Monika PANDEY*
Mastercard
Pune, India
monika.pandey@mastercard.com

ABSTRACT

We present AVA, an agentic, hands-free voice assistant that provides interactive, contextual guidance in real-world environments such as makerspaces. AVA combines large language models with retrieval-augmented generation over environment-specific knowledge (e.g., manuals and safety procedures) and an explicit agent decision/control graph to infer user intent, estimate confidence in retrieved information, and choose dialogue actions such as answering directly, asking clarifying questions, or guiding exploration. The system supports coherent multi-turn interaction via lightweight session memory and augments spoken responses with dynamically generated QR codes linking to authoritative resources. AVA demonstrates real-time, situated assistance for tool discovery, safe operation, and project ideation, highlighting the value of agent-based architectures for scalable, consistent support in evolving physical spaces.

KEYWORDS

Voice Assistant; Agentic Architecture; Situated Interaction

ACM Reference Format:

Donghao HUANG* and Monika PANDEY*. 2026. An Agentic Voice-Based Assistant for Interactive Conversation and Guidance in Real-World Environments: Demonstration Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/RGWK4224

1 INTRODUCTION

Physical environments like innovation labs, shared workspaces, and educational makerspaces increasingly depend on diverse tools, procedures, and safety rules. Yet general-purpose voice assistants are not designed for these settings: they lack awareness of local inventories, operational constraints, and context-specific documentation, which limits their usefulness for situated, task-oriented help. At the same time, relying on human staff does not scale—availability varies, expertise may be uneven, and support quality can be inconsistent as toolsets evolve and user questions remain open-ended. Makerspaces illustrate this challenge especially well, since users often struggle to discover the right tools, understand safe operation, and connect documentation across fabrication, robotics, and electronics workflows. To address this gap, we demonstrate[1] an agentic, hands-free voice assistant that combines large language models [5] with retrieval over environment-specific knowledge [4] and an explicit agent control framework [6, 8]. The agent infers user intent, estimates confidence in retrieved information, and selects dialogue actions such as answering directly, asking clarifying questions, or guiding exploration. It supports coherent multi-turn interaction through lightweight session memory and augments spoken guidance with dynamically generated QR codes linking to authoritative manuals and resources. Overall, the system highlights how agent-based architectures can provide scalable, context-grounded assistance for tool discovery, safe operation, and creative ideation beyond what general-purpose voice technologies can offer.

*Equal contribution.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/RGWK4224

2 SYSTEM OVERVIEW

Application Scenario. The system is deployed as a walk-up kiosk in a Makerspace where users interact via natural spoken queries (e.g., *"What can I do here?"* or *"How do I use the robotic dog?"*). The assistant interprets intent, retrieves environment-specific knowledge, and delivers hands-free spoken guidance, augmented by dynamically generated QR codes linking to manuals and safety documentation. While demonstrated in a Makerspace, the design generalizes to innovation labs, educational facilities, and other instrumented physical environments.

Architecture. As shown in Figure 1, the system adopts a modular, agent-centric architecture that separates interaction handling, agent reasoning, knowledge access, and output grounding.

- **Speech Interface**: Provides bidirectional voice interaction via speech-to-text and text-to-speech [9]. Continuous listening with silence-based segmentation enables hands-free, walk-up use.
- **Agent Core**: Implements autonomous decision-making as an explicit state-based control graph progressing through perception, intent inference, deliberation, and response generation. Utterances are classified into general queries, tool-specific requests, or session termination signals. Lightweight session memory captures dialogue context and salient entities. A key capability is confidence-aware deliberation: the agent assesses information sufficiency and autonomously selects its next action—direct response, clarification, or redirection—outside the language model, ensuring action selection is governed by agent state rather than model output alone.

[1]Demo video available at https://vimeo.com/1153194189

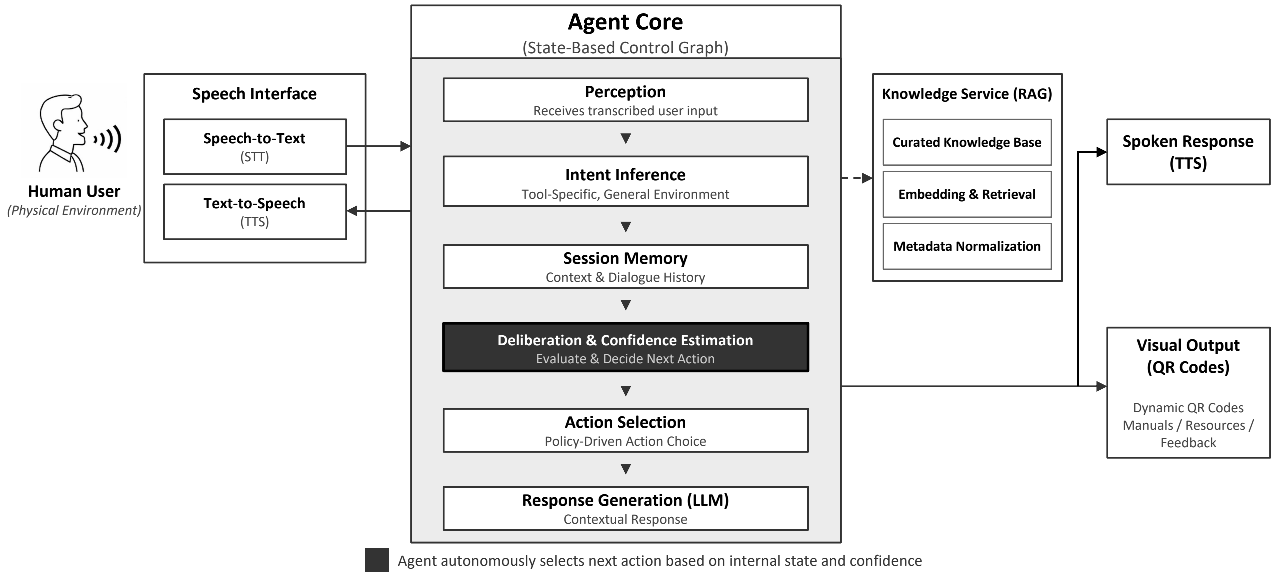

Figure 1: Agentic architecture of the voice-based assistant.

- **Knowledge Service**: Implements retrieval-augmented generation (RAG) over a curated knowledge base of tool manuals, procedures, and safety guidelines. Retrieved documents are annotated with metadata and injected into the LLM context together with session memory to enable grounded responses.
- **Visual Output**: Dynamically renders contextual QR codes linking to referenced tools or resources, providing seamless access to detailed documentation without interrupting the interaction flow.

Agentic Behavior. The system exhibits agentic autonomy through a confidence-aware decision policy. After retrieving candidate knowledge via the RAG service, the agent evaluates evidence sufficiency and maintains an internal confidence estimate. When confidence is high, it produces a grounded response; when low or underspecified, it autonomously initiates clarifying questions. This behavior is realized through state transitions in the agent control graph rather than prompt engineering alone [1], enabling context-sensitive action selection driven by agent state and environmental knowledge [7].

Relation to BDI and Neurosymbolic Agent Architectures. Recent work has explored combining the principled deliberation of Belief–Desire–Intention (BDI) agents with the natural language capabilities of large language models. ChatBDI [2], for instance, adds conversational interfaces to existing BDI agents by translating natural language into KQML performatives, endowing BDI agents with fluent communication while preserving their intentional reasoning. Similarly, other BDI+LLM approaches focus on plan generation [3] or goal-directed autonomy within symbolic agent programming frameworks. AVA's agent control graph shares the BDI commitment to explicit deliberation—intent inference, confidence assessment, and action selection are governed by structured state transitions rather than end-to-end prompting—but differs in two respects. First, AVA targets real-time, situated voice interaction in physical environments, where latency and hands-free usability impose tight constraints on the deliberation cycle. Second, rather than coupling to a full BDI interpreter, AVA adopts a lightweight decision graph that orchestrates an LLM together with RAG retrieval and session memory, balancing interpretability with the flexibility needed for open-ended user queries in evolving physical spaces.

AVA Interaction Flow. The live demonstration illustrates an end-to-end user journey, showing how the agent supports exploration, guidance, and feedback in a physical environment:

(1) **Walk-Up Engagement.** A visitor initiates interaction through natural speech at a kiosk. The system automatically establishes a session, enabling continuous, hands-free use without prior setup.

(2) **Exploratory Discovery.** Users ask open-ended questions such as *"What can I do here?"*. The agent interprets exploratory intent and provides a concise overview of available tools and activities.

(3) **Targeted Guidance.** For tool- or process-specific queries, the agent retrieves environment-specific knowledge and delivers short spoken guidance, emphasizing correct usage and safety.

(4) **Contextual Deep-Dive.** When additional detail is relevant, the system displays dynamically generated QR codes linking to official manuals or safety documentation for continued learning.

(5) **Adaptive Multi-Turn Dialogue.** The agent maintains lightweight session memory, supporting coherent follow-up questions and adaptive refinement of responses across multiple turns.

(6) **Clarification Under Uncertainty.** If a request is ambiguous or insufficiently specified, the agent autonomously asks clarifying questions, demonstrating confidence-aware decision making.

(7) **Project Ideation Support.** Beyond procedural assistance, the agent suggests potential projects and tool combinations, encouraging creative exploration aligned with local resources.

(8) **Feedback and Session Closure.** At session end, the agent invites user feedback and presents a QR code linking to a feedback form, supporting continuous system improvement.

3 ACKNOWLEDGMENTS

This work was supported by Mastercard Foundry R&D. The authors thank Varuna Ektare from the Mastercard Foundry Pune R&D team for her assistance with the program management. This work was also supported by the EngD Program at Singapore Management University.

REFERENCES

[1] Louise Dennis, Michael Fisher, and Marija Slavkovik. 2016. Formal Verification of Autonomous Systems. *AI Magazine* (2016).
[2] Andrea Gatti. 2025. ChatBDI: Think BDI, Talk LLM. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).*
[3] Alexandre Yukio Ichida, Felipe Meneguzzi, and Rafael C. Cardoso. 2024. BDI Agents in Natural Language Environments. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS).*
[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS.*
[5] OpenAI. 2023. *GPT-4 Technical Report.* Technical Report arXiv:2303.08774.
[6] Stuart Russell and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
[7] Terry Winograd and Fernando Flores. 1986. *Understanding Computers and Cognition.* Addison-Wesley.
[8] Michael Wooldridge. 2009. *An Introduction to MultiAgent Systems* (2nd ed.). Wiley.
[9] Xiaoxia Yu and John Hansen. 2019. Speech Recognition in Human–Machine Interaction: A Survey. *IEEE Signal Processing Magazine* (2019).