Bruna: A Real-Time Multimodal Voice Agent with Hybrid Reasoning

Evandro Fonseca

Blip

evandro.fonseca@blip.ai

Abstract

This paper describes Bruna, a data-centric smart voice assistant powered by multiple Large Language Models designed to support Stilingue and Blip products. Our architecture provides an enriched conversational experience, delivering strategic insights in real-time.

1 Introduction

Currently, data analysts and decision-makers frequently face cognitive overload in their daily workflows. Although (Fonseca et al., 2024) propose a copilot capable of providing suggestions based on conversational context, optimizing the experience for customer service agents we identify a gap regarding the processing of dynamic, multimodal contexts. This limitation motivates us to explore novel approaches designed to deliver a more immersive analytical experience.

To address this challenge, we present Bruna, a voice assistant designed to streamline data interaction by processing audio and visual inputs simultaneously. Unlike text only systems, Bruna enables users to interact naturally with complex data representations.

Our architecture leverages the Model Context Protocol (MCP) (Hou et al., 2025) to dynamically connect the agent with external tools. To ensure low latency and fluid interaction, we utilize an asynchronous pipeline based on Server Sent Events (SSE) and the Azure OpenAI Realtime API[1]. This allows the agent to perceive visual content, such as charts on a screen and cross reference it with live market data on demand, providing immediate, hands-free strategic insights.

2 Architecture

Our architecture is designed to be modular and agnostic regarding the underlying Large Language Models (LLMs). In our implementation, we adopted a composite strategy that leverages the strengths of multiple models. For the conversational interface, we utilized Azure OpenAI Realtime GPT-4o to ensure low-latency verbal interaction. However, in our experiments, we observed that models optimized for real-time audio processing often lack efficient reasoning capabilities for complex analytical tasks. Therefore, seeking support from more robust text-based models is a fundamental part of our strategy. To address this, the system offloads complex data processing and reasoning tasks to Gemini 2.5 Flash-Lite(Comanici et al., 2025), ensuring that the agent remains both responsive and intellectually capable.

When developing applications that integrate real-time audio with external execution tools, interoperability is a major challenge. At the time of our agent's conception, we did not find foundation models with native support for the Model Context Protocol (MCP). Consequently, it was necessary to construct a custom MCP Adapter. This component acts as a bridge, standardizing the communication between the proprietary real-time API and any standard MCP server. This allows our agent to connect seamlessly with diverse external tools, such as Radar Stilingue or vector databases, regardless of the underlying model's native capabilities.

Recent work by Mileff (Mileff, 2025) addresses the latency challenge in voice agents by proposing a parallelized architecture that orchestrates WebSockets and multi-threaded Text-to-Speech (TTS) services. While Mileff's approach effectively minimizes the delay between text generation and audio synthesis through a segmented pipeline, our architecture differs by adopting a native multimodal stream approach. Instead of optimizing the serialization of text-to-audio, we integrate the tool execution layer directly into the context loop. In this way, Bruna does not merely read generated text faster; she suspends the audio stream to "think"

[1]available in https://learn.microsoft.com/pt-br/azure/ai-foundry/openai/realtime-audio-quickstart

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 11–13
April 13-16, 2026

(process data via Gemini/MCP) only when deep reasoning is required, resuming the conversation with validated insights.

Figure 1 shows the complete pipeline. When a session is initiated, the MCP Adapter registers the available tools based on the user's permissions. Throughout the interaction, the audio input is processed by the Realtime model. If a complex intent is detected (e.g., "Analyze this dashboard"), the adapter intercepts the request, routes the visual and textual context to Gemini Flash 2.5 for reasoning, and returns the synthesized insight to the audio model for verbal delivery. This hybrid approach ensures that the system maintains the fluidity of a real-time voice agent while possessing the analytical depth of a large-scale text model.

3 Interface and Use Cases

Typically, extracting actionable intelligence from social listening platforms requires navigating through complex dashboards, applying multiple filters, and manually interpreting high-dimensional charts. This process can be time-consuming and cognitively demanding for decision-makers who need immediate strategic answers. Considering this friction, the Voice Agent Controller interface was designed not merely as a replacement for visual dashboards, but as a conversational abstraction layer that simplifies access to complex data streams.

As illustrated[2] in Figure 2, the web interface is designed to be minimalist, prioritizing the audio-visual interaction stream. The layout consists of session controls (start/stop), media toggles (camera/microphone), and a granular real-time event log. This log serves a critical role in system transparency and explainability, displaying *Server-Sent Events* (SSE) regarding connection status, tool execution steps, and token consumption metrics. This allows the user to monitor the agent's "reasoning" process in real-time, validating that the correct external tools are being invoked before an audio response is synthesized.

3.1 Case Study: Strategic Marketing Analysis

To validate the effectiveness of the multimodal architecture and the reasoning capabilities of the hybrid model strategy, we conducted a case study

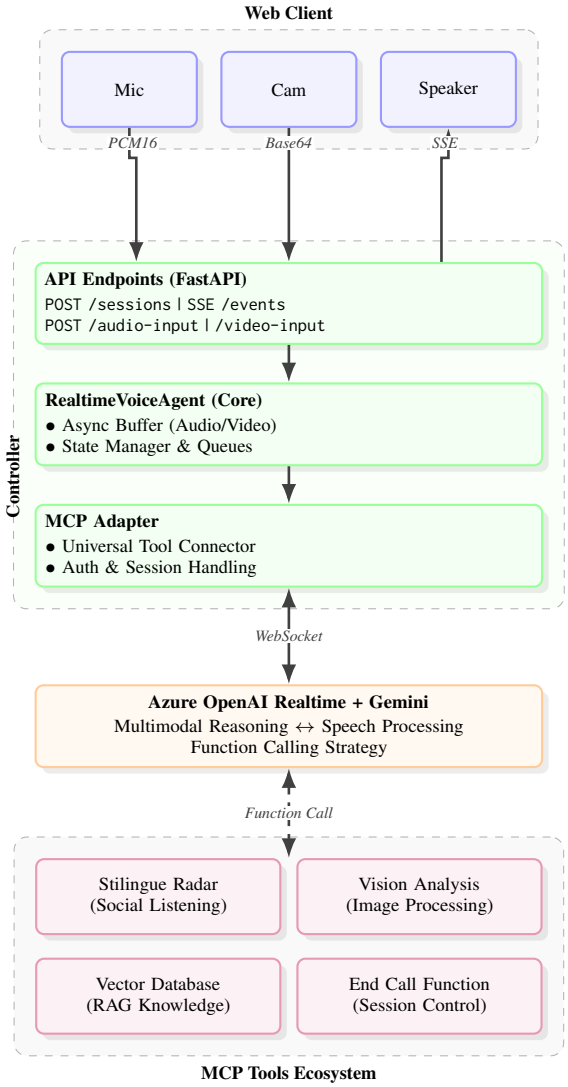

Figure 1: Architecture of the Voice Agent Controller. The system buffers multimodal inputs (audio/video) and orchestrates dynamic tool execution via the Model Context Protocol (MCP) Adapter. This layer facilitates Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and real-time data access to ground the model's reasoning before interacting with the Azure OpenAI Realtime API.

[2]Bruna's web interface is available to test in: https://secure-backend-api.stilingue.com.br/rd-envision-mcp-server/prod/voice-agent/test

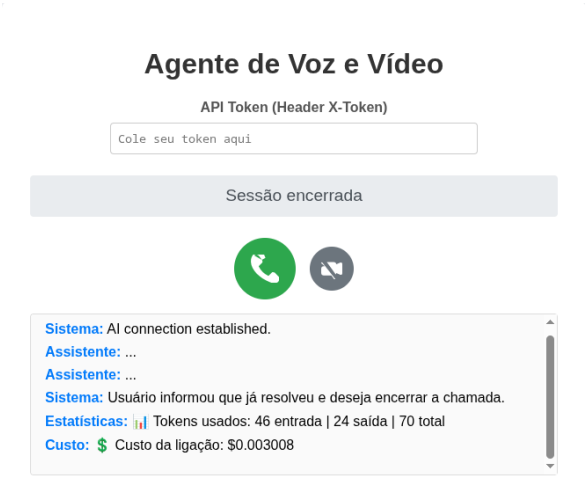

Figure 2: The Voice Agent Controller interface. The central panel manages the WebRTC media streams, while the bottom log provides real-time feedback on system events, tool execution, and token usage costs.

simulating a strategic planning session. In this scenario, the user acts as a marketing manager for a major fashion retail giant, aiming to structure a campaign for the Carnival holiday.

The interaction demonstrates the system's ability to convert raw data into strategic differentiation:

1. **Contextual Query:** The user requests a benchmark analysis of competitors in the apparel sector to guide a Carnival campaign strategy.

2. **Autonomous Execution:** The agent identifies the need for real-time data and triggers the social listening tool via MCP to query live sentiment and trending topics.

3. **Insight Generation:** The analysis detects a market opportunity: high positive engagement for "promotions" but significant negative sentiment regarding "delivery delays" among competitors.

4. **Strategic & Creative Output:** Bruna synthesizes a strategy focusing on reliability to counter competitor weaknesses and proposes campaign names like *"Carnaval sem Perrengue"* (Hassle-free Carnival).

This case study highlights the system's capacity to function as a "proactive copilot." By abstracting the complexity of database queries and sentiment analysis into a natural conversation, the interface allows users to focus on high-level decision-making rather than data mining.

4 Conclusion

In this paper, we presented Bruna, a multimodal voice agent that leverages the Model Context Protocol (MCP) and hybrid reasoning to generate real-time strategic insights. We detailed our architecture designed for low latency, showing how it reduces the cognitive load required to interpret complex data and maximizes the efficiency of decision-making processes. By maintaining conversational fluidity while accessing external tools, the system bridges the gap between static dashboards and active intelligence.

As further work, we intend to integrate the agent into the workflows of the Stilingue and Blip platforms. We also plan to conduct quantitative and qualitative studies to measure how the agent optimizes user time and reduces the "time-to-insight" compared to traditional visual dashboard navigation.

References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Evandro B Fonseca, Tayane Soares, Dyovana Baptista, Rogers Damas, and Lucas Avanço. 2024. Blip copilot: a smart conversational assistant. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 194–196.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Péter Mileff. 2025. Real-time, low audio latency based ai-powered application architecture design. *Production Systems and Information Engineering*, 13(1):46–63.