Designing Explainable Conversational Agentic Systems for Guaraní Speakers

Samantha Adorno[1], **Akshata Kishore Moharir**[2], **Ratna Kandala**[1]

[1]University of Kansas, Lawrence, Kansas, USA
[2]Independent Researcher, Sherwood, Oregon, USA
samantha.adorno00@gmail.com, akshatankishore5@gmail.com, ratnanirupama@gmail.com

Abstract

Artificial intelligence systems are often presented as universal, yet their interaction paradigms remain predominantly text-first, limiting alignment with primarily oral languages and communicative practices. Using Guaraní, an official and widely spoken language of Paraguay, as a motivating case, this work examines how language support risks remaining symbolic when spoken interaction is reduced to a speech-to-text interface. We explore an oral-first, multi-agent framing in which turn-taking, repair, shared context, and governance are treated as core components of interaction rather than peripheral features. By separating language understanding from conversation state and permission mechanisms, the architecture makes conversational structure and control explicit, enabling reasoning over interaction dynamics rather than isolated commands.

Introduction

Most AI systems and everyday interaction with machines remain oriented around text input, such as keyboards, menus, and form-like interfaces. Voice features are increasingly common, but in many deployments they function primarily as a spoken front-end to a text-first pipeline (transcribe, parse, respond) rather than as sustained conversation with turn-taking, clarification, and repair. Such systems frequently struggle with interruptions, response timing, and turn coordination (Skantze 2021), and breakdown handling often places the burden on users to rephrase or restart (Alghamdi, Halvey, and Nicol 2024).

Conversation, however, is coordinated action. People manage understanding through grounding and shared context, using clarification and repair to maintain alignment (Clark and Brennan 1991). When systems cannot manage turn-taking and repair, voice interaction collapses into brittle command-and-control behavior, even when speech recognition and synthesis are available (Skantze 2021; Alghamdi, Halvey, and Nicol 2024). These limitations disproportionately affect low-resource languages, many of which are primarily oral and lack extensive written or dialogue resources (Turin 2012).

We use Guaraní as a motivating case. Guaraní is one of Paraguay's two official languages and is widely used in daily life (Organization of American States (OAS) 1992). Interface design in Paraguay operates within a context of *diglossia*: a stable separation of language functions across domains, with a "High" (H) code used in formal and written contexts and a "Low" (L) code used in everyday speech (Ferguson 1959; Fishman 1967). In practice, Guaraní predominates in oral and community settings, while Spanish dominates literacy, bureaucracy, and public-facing written systems (Ito 2012). Because most digital interaction is mediated through written interfaces, systems implicitly privilege Spanish literacy, creating a *domain mismatch* between everyday reasoning and digital infrastructure. Internet use in Paraguay has grown substantially, rising from 61.1% in 2017 to 81.6% in 2024 among those aged 10+ (Instituto Nacional de Estadística (INE), Paraguay 2025), yet increased connectivity has not translated into broader Guaraní integration in digital services.

Rather than adapting oral languages to text-centric systems, oral interaction should be treated as a first-class design starting point. Orality-grounded HCI argues that assumptions imported from literate settings can fail in oral cultures, where knowledge organization relies on narrative structure, repetition, and socially distributed memory (Sherwani et al. 2009). The challenge is therefore not only recognition or translation, but supporting multi-turn interaction, shared context, and repair in the language users actually speak.

Proposed Architecture: An Oral-First Multi-Agent System

To address this, we propose an *oral-first* architecture that separates language understanding, conversation state, action execution, and governance into interacting agents. A Multi-Agent System (MAS) framework enables specialization while keeping interfaces explicit and inspectable (Becker 2024), and turn-taking and repair require dedicated state tracking distinct from generation or execution (Sapkota et al. 2025). The system orchestrates six specialized agents:

- **Speech Interface Agent:** Performs Voice Activity Detection and turn segmentation, using pause duration and timing cues to distinguish floor-holding from turn completion (Skantze 2021; Stivers et al. 2009). This is critical for Guaraní, where brief pauses and glottal stops (*puso*)

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

can occur within words and should not be misinterpreted as turn endings (Estigarribia 2020).

- **Guaraní Understanding Agent:** Interprets Guaraní and Jopará speech by mapping utterances to structured intents and slots (e.g., `PLAY_MUSIC`, `OPEN_TAB`). It is designed to operate on community-verified speech data to capture culturally specific references and loanword usage.
- **Conversation State Agent:** Maintains dialogue memory across turns, enabling resolution of ellipsis and implicit references.
- **Permission & Governance Agent:** Operates as a sovereign mediation layer between interpretation and execution. It evaluates requested actions against user consent settings and community-defined privacy norms (e.g., default "do not store audio") before allowing downstream execution.
- **Response Agent:** Generates conversational responses grounded in the current dialogue state, which are checked by the Permission & Governance Agent before being delivered to the user.
- **Action Agents:** Domain-specific executors (e.g., Media or Browser agents) responsible for carrying out approved actions through external tools or APIs.

Table 1 illustrates how the agents coordinate across a two-turn exchange, including implicit reference resolution in the second turn.

User (Guaraní)	**System Action & Agent Logic**
"Che ahenduse purahei" (I want to listen to music)	1. **Speech Agent** captures audio. 2. **Understanding Agent** identifies intent: `PLAY_MUSIC`. 3. **Permission Agent** verifies "Music" is a safe category; approves. 4. **Media Agent** selects a popular playlist and starts playback.
Response Agent	*"Oĩporã"* (Ok / It is good)
"Nda che gustái" (No, I don't like [this])	1. **Understanding Agent** detects negative sentiment/rejection. 2. **Conversation State Agent** resolves the implicit object: "this" = *current song*. Updates intent to `SKIP`. 3. **Media Agent** executes `NEXT_TRACK`.

Table 1: Multi-turn interaction demonstrating context resolution and repair.

A central question for the Guaraní Understanding Agent is how to obtain authentic, community-verified speech rather than relying on synthetic translations. Mozilla Common Voice (Guaraní, `gn`) provides a large, crowdsourced open speech resource that supports multilingual acoustic modeling (Ardila et al. 2020). Community-led efforts such as *Aikuaa*, which organizes collaborative "mingas" to collect and validate Guaraní voice data, further capture *Jopará* usage and conversational variation often absent from formal or archival datasets (JournalismAI 2025).

Implementation Strategy

The proposed architecture is hybrid by design. Each agent would be assigned the lightest sufficient mechanism: large language models for components where flexible interpretation of speech and dialogue is required, and deterministic logic where auditability, low latency, or safety dominate.

The Speech Interface Agent would be rule-based, with VAD thresholds and pause-duration heuristics tuned for Guaraní prosody, including the within-word *puso*. The Guaraní Understanding Agent would use a fine-tuned multilingual model such as NLLB-200 or a Whisper-derived encoder, mapping speech to intent-slot structures through a constrained output schema rather than free-form generation. The Conversation State Agent would be a structured store of dialogue state (active intent, salient referents, last system action), updated by deterministic rules and queried by other agents. The Response Agent would be LLM-driven but conditioned on the state store and restricted to short, register-appropriate replies. The Permission & Governance Agent and Action Agents would be rule-based, with policy decisions and tool calls expressed as explicit, inspectable rules.

This separation is intended to keep the components most likely to fail silently, those handling consent, action execution, and turn segmentation, outside the generative path. It also supports component-level ablation, since an oracle-model substitute can replace any LLM-based agent without altering the surrounding system.

Operationalizing Governance and Consent

The Permission & Governance Agent mediates between interpretation and execution. Each candidate action carries an action type, target, and audio-retention flag, evaluated against three layers: (1) per-session user consent set at the start of interaction, (2) persistent user-level defaults, and (3) community-defined norms (e.g., audio is never stored by default; sensitive categories such as health or finance require explicit per-turn confirmation). The agent returns approval, denial, or a confirmation request to the orchestrator before any Action Agent runs. The following exchange illustrates a sensitive-category interaction:

> *User: "Embohasa pe pohã repykue"* (Send the medication payment)
> *Permission Agent:* category = `FINANCE_HEALTH`, retention = off, requires confirmation.
> *Response Agent: "Reipota piko amondo?"* (Do you want me to send it?)
> *User: "Hee"* (Yes) → Action executed; audio not retained.

This makes consent observable per turn and inspectable after the fact, rather than buried in opaque settings.

Data Collection Plan

Multi-turn Guaraní conversational audio remains scarce. We outline a three-stage protocol building on existing community infrastructure, to be carried out through planned in-country fieldwork in Paraguay in partnership with local Guaraní-speaking communities and organizations such as *Aikuaa*. (1) *Seed prompts*: a small set of task-oriented scenarios (media control, browsing, payments) co-designed

with Guaraní speakers rather than translated from English templates. (2) *Minga-style recording*: paired and small-group sessions modeled on the Aikuaa workflow (JournalismAI 2025), capturing natural turn-taking, overlaps, and repair rather than read-aloud single utterances. (3) *Community validation*: speakers other than the recorders annotate intents, resolve referents, and flag culturally specific references, with explicit consent for each retention decision per the CARE principles (Carroll et al. 2020). Annotation focuses on multi-turn structure: turn boundaries, repair type (clarification, correction, abandonment), and referent chains. Mozilla Common Voice Guaraní (Ardila et al. 2020) would supply acoustic pretraining; the minga corpus would supply the multi-turn structure absent from existing resources.

Evaluation Criteria

Evaluating an oral-first architecture requires metrics beyond accuracy that capture context, sentiment, and privacy in multi-turn interaction. We consider four dimensions of conversational success:

- **Task Success Rate (TSR):** The percentage of multi-turn goals completed successfully, capturing dialogue coherence and evolving intent interpretation across turns.
- **Repair Success Rate:** How often the system recovers from errors, such as misheard words or misunderstood intents, without requiring users to restart their task.
- **Perceived Sovereignty:** A qualitative metric evaluating user confidence that audio is not stored or reused without consent, requiring community-centered, ethnographic evaluation methods.
- **Latency:** Whether system response timing aligns with Guaraní conversational tempo, avoiding premature interruptions that violate turn-taking norms and prolonged silences that disrupt conversational flow.

A planned ablation isolates the contribution of each agent. Removing the Conversation State Agent should degrade Repair Success on second-turn implicit-reference cases. Replacing the Understanding Agent with an oracle transcript-to-intent mapping bounds the headroom available from acoustic improvements. Disabling the Permission & Governance Agent should leave TSR largely unchanged while substantially reducing Perceived Sovereignty. This separation ties each empirical claim to a specific component.

Generalization Beyond Guaraní

The architecture is shaped by Guaraní, but its design commitments transfer to other primarily oral and low-resource settings. The Speech Interface, Understanding, and Response agents are language-specific and require local adaptation. The Conversation State and Action Agents are largely language-agnostic and can be reused. The Permission & Governance Agent is a structural commitment rather than a portable artifact: the separation of consent and execution transfers, but the policies enforced inside it must be co-defined with each community, since different Indigenous and oral-language communities hold different norms around recording, retention, and data sharing. Adapting the system to another oral or diglossic language therefore amounts to retraining the Understanding and Response agents on community-validated data, retuning VAD for that language's prosody, and re-eliciting governance policies in partnership with local speakers. Cross-language transfer remains an open empirical question and a direction for future work.

Discussions and Limitations

A persistent challenge for Guaraní language technology is the gap between institutional standardization and everyday speech. Although Guaraní is an official language supported by bodies such as the Academia de la Lengua Guaraní under the Ley de Lenguas (Secretaría de Políticas Lingüísticas (Paraguay) 2010; Secretaría de Políticas Lingüísticas), daily use frequently involves *jopará* and code-switching (Mortimer 2006; Estigarribia 2015; Kellert and Tyagi 2025). Oral-first systems should therefore treat variation as expected input and prioritize communicative intent over enforcing a single normative register (Mortimer 2006; Kellert and Tyagi 2025).

Beyond general data scarcity, oral-first systems lack conversational audio capturing turn-taking, repair, and shared context. For Guaraní, existing corpora and scraping-based methods mainly support text-based evaluation rather than spontaneous multi-turn speech (Chiruzzo et al. 2022; Góngora, Giossa, and Chiruzzo 2021). Future efforts should prioritize community-led collection of conversational data (JournalismAI 2025).

Oral interfaces raise governance challenges because speech data is inherently identifiable and easily repurposed. Indigenous data governance frameworks emphasize community benefit, control, and accountability (Carroll et al. 2020). This motivates separating execution from a dedicated permission and privacy layer that mediates consent and data retention, including explicit "do not store audio" defaults. Such separation supports ethical commitments while increasing perceived user control and trust in multi-turn interaction (Carroll et al. 2020; Alghamdi, Halvey, and Nicol 2024).

Conclusion

This work contributes to culturally grounded AI by treating conversation as a core computational structure rather than an interface layer. When interaction models fail to reflect how language is practiced, language support risks remaining symbolic rather than operational. Using Guaraní as a motivating case, we outline a multi-agent architecture that elevates turn-taking, repair, and shared context to first-class system components. By separating language understanding from explicit permission and governance mechanisms, the architecture makes conversational reasoning, control, and accountability inspectable and modular. More broadly, the work argues that equitable and aligned AI systems must reason over conversation as it is lived, particularly in oral and low-resource settings, rather than adapting those settings to inherited text-centric paradigms.

References

Alghamdi, E.; Halvey, M.; and Nicol, E. 2024. System and User Strategies to Repair Conversational Breakdowns of Spoken Dialogue Systems: A Scoping Review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705113.

Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Becker, J. 2024. Multi-Agent Large Language Models for Conversational Task-Solving. arXiv preprint arXiv:2410.22932v1.

Carroll, S. R.; et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*.

Chiruzzo, L.; Góngora, S.; Alvarez, A.; Giménez-Lugo, G.; Agüero-Torales, M.; and Rodríguez, Y. 2022. Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2098–2107. Marseille, France: European Language Resources Association.

Clark, H. H.; and Brennan, S. E. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*. American Psychological Association.

Estigarribia, B. 2015. Jopará and Guaraní in Paraguay (discussion of contact and mixed speech).

Estigarribia, B. 2020. *A Grammar of Paraguayan Guaraní*. UCL Press.

Ferguson, C. A. 1959. Diglossia. *Word*, 15(2): 325–340.

Fishman, J. A. 1967. Bilingualism with and without Diglossia; Diglossia with and without Bilingualism. *Journal of Social Issues*, 23(2): 29–38.

Góngora, S.; Giossa, N.; and Chiruzzo, L. 2021. Experiments on a Guarani Corpus of News and Social Media. In Mager, M.; Oncevay, A.; Rios, A.; Ruiz, I. V. M.; Palmer, A.; Neubig, G.; and Kann, K., eds., *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 153–158. Online: Association for Computational Linguistics.

Instituto Nacional de Estadística (INE), Paraguay. 2025. 8 de cada 10 personas utiliza internet en Paraguay (EPH 2017–2024).

Ito, H. 2012. With Spanish, Guaraní lives: a sociolinguistic analysis of bilingual education in Paraguay. *Multilingual Education*, 2(1): 6.

JournalismAI. 2025. Guarani AI: When building language tech means building community.

Kellert, O.; and Tyagi, N. 2025. Where and How Do Languages Mix? A Study of Spanish-Guaraní Code-Switching in Paraguay. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics.

Mortimer, K. 2006. Guaraní Académico or Jopará? Educator Perspectives and Ideological Debate in Paraguayan Bilingual Education.

Organization of American States (OAS). 1992. Paraguay's Constitution of 1992 with Amendments through 2011. PDF.

Sapkota, S.; et al. 2025. Multi-Party Conversational Agents: A Survey. arXiv preprint arXiv:2505.18845v1.

Secretaría de Políticas Lingüísticas. ???? Academia de la Lengua Guaraní. https://spl.gov.py/es/academia-de-la-lengua-guarani/.

Secretaría de Políticas Lingüísticas (Paraguay). 2010. Ley Nº 4251/2010: Ley de Lenguas (texto bilingüe). PDF.

Sherwani, J.; Ali, N.; Penstein Rosé, C.; and Rosenfeld, R. 2009. Orality-Grounded HCID: Understanding the Oral User. *Information Technologies & International Development*, 5(4): 37–49.

Skantze, G. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67: 101178.

Stivers, T.; Enfield, N. J.; Brown, P.; Englert, C.; Hayashi, M.; Heinemann, T.; Hoymann, G.; Rossano, F.; de Ruiter, J. P.; Yoon, K.-E.; and Levinson, S. C. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26): 10587–10592.

Turin, M. 2012. Voices of vanishing worlds: Endangered languages, orality, and cognition. *Análise Social*, 205(47).