An Oral-first Interactive Agentic System for Guaraní Speakers

Samantha Adorno, Akshata Kishore Moharir, Ratna Kandala
University of Kansas, Independent Researcher, University of Kansas
samantha.adorno00@gmail.com, akshatankinshore5@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 the conversation state and permission mechanisms, the architecture makes conversational structure and control explicit, enabling reasoning over interaction dynamics rather than isolated commands. Framing conversational coordination as a cognitively motivated reasoning problem over shared state connects insights from human dialogue to the design of AI systems that are more interpretable and responsive in oral and low-resource settings.

Keywords: Conversational AI, Oral-first Interaction, Low-resource Languages, Diglossia, Guaraní, Culturally Grounded AI, Multi-Agent Systems, Indigenous Data Governance

1. 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. Voice assistants such as Amazon Alexa illustrate this interaction style: wake word, short request, single response. 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 et al., 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). Turn-taking is central to this coordination, with measurable cross-cultural variation in timing (Stivers et al., 2009). 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 et al., 2024). These limitations disproportionately affect low-resource languages, many of which are primarily oral and lack extensive written or dialogue resources (Turin, 2012).

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). Although recent projects have improved speech recognition for underrepresented speech communities, such as Aboriginal English in Australia (Hutchinson, 2025; The University of Western Australia, 2025), most systems remain structured around text-first interaction assumptions and lack explicit mechanisms for dialogue state tracking, repair, and consent. Prior work on Guaraní speech recognition has similarly highlighted the challenges of low-resource settings: competition results covering Guaraní alongside other Indigenous languages of the Americas show poor performance largely attributable to data scarcity (Ebrahimi et al., 2023), and targeted efforts to fine-tune models such as Whisper specifically for Guaraní further underscore the gap between off-the-shelf multilingual models and the needs of the language (Acevedo Zarza et al., 2024).

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). While most speakers regularly use Guaraní, Spanish, or both (Instituto Nacional de Estadística (INE), Paraguay, 2024), digital systems continue to privilege Spanish for interaction and disambiguation. The challenge is therefore not only recognition or translation, but supporting multi-turn interaction, shared context, and repair in the language users actually speak.

To address this, we propose an oral-first assistant architecture that separates language understanding, conversation state, action execution, and governance into interacting agents. A Multi-Agent

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 9–14

System (MAS) framework enables specialization while keeping interfaces explicit and inspectable. Prior work shows that decomposing conversational task-solving into specialized agents improves performance relative to monolithic models (Becker, 2024), and that turn-taking and repair require dedicated state tracking distinct from generation or execution (Chen et al., 2025).

2. Case Study: Guaraní in Paraguay

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-menus, forms, and error messages-systems implicitly privilege Spanish literacy. Even when Guaraní is supported, interaction often defaults to Spanish at moments requiring verification or correction, creating a *domain mismatch* between everyday reasoning and digital infrastructure (Ito, 2012). 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). However, increased connectivity has not translated into broader Guaraní integration in digital services. Mainstream platforms remain optimized for Spanish literacy and symbolic input, meaning language support often remains nominal rather than functionally embedded in interaction. This tension between everyday oral practice and text-centric digital systems motivates the need for interaction models that treat spoken coordination, repair, and shared context as primary rather than secondary design considerations.

3. Proposed Architecture: An Oral-First Multi-Agent System

Figure 1 illustrates the proposed system architecture, showing how Guaraní voice input is routed through a pipeline of specialized agents before any action is executed.

We propose an *oral-first* architecture that treats speech as the primary interaction modality and makes multi-turn context, repair, and governance explicit. 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*) can occur within words and should not be misinterpreted as turn endings (Estigarribia, 2020).

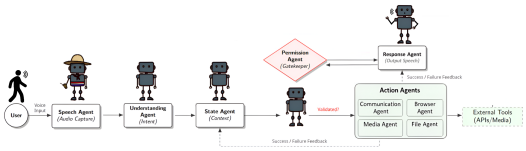

Figure 1: Proposed Oral-First Multi-Agent Architecture. Voice input is routed through speech capture, intent understanding, conversation state tracking, and a permission and governance layer before tool execution. The Speech Agent (far left) is depicted in traditional Paraguayan attire to symbolize the system's grounding in local cultural identity.

- **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 a sample multi-turn exchange, demonstrating how the agents coordinate to handle context resolution and repair across turns.

3.1. Training Data and Community Alignment

A central question for the Guaraní Understanding Agent is how to obtain authentic, community-verified speech rather than relying on synthetic

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	*"Oĩ porã"* (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.

translations. Two initiatives illustrate the available resources. Mozilla Common Voice (Guaraní, `gn`) provides a crowdsourced open speech resource that supports multilingual acoustic modeling (Ardila et al., 2020), and community-led efforts such as *Aikuaa*, organized by *El Surtidor* through collaborative "mingas," further capture *Jopará* usage and conversational variation often absent from formal datasets (JournalismAI, 2025). However, both resources consist primarily of read-aloud sentences and neither contains the task-oriented or multi-turn utterances this system requires.

Addressing this gap requires a dedicated collection effort. We plan to organize community recording sessions modeled on the Aikuaa minga format, in which native speakers are prompted with realistic task scenarios (e.g., controlling media, opening a browser tab, asking follow-up questions) and asked to respond naturally in Guaraní or Jopará. Critically, prompts will be presented orally rather than as written text, to avoid literacy-based anchoring of responses. Sessions will be designed to capture: (1) *task-oriented utterances* covering the intent categories the system supports; (2) *multi-turn exchanges* that include ellipsis, implicit reference, and repair; and (3) *code-switching patterns* reflecting everyday mixed-language use. All recordings will be collected under explicit community consent protocols aligned with the Permission & Governance Agent's design, with speakers retaining control over storage and reuse of their voice data.

3.2. 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):** Measures the percentage of multi-turn goals completed successfully, capturing whether the Conversation State Agent maintains dialogue coherence and whether the Understanding Agent correctly interprets evolving intents across turns.
- **Repair Success Rate:** Measures conversational resilience by evaluating how often the system recovers from errors, such as misheard words or misunderstood intents, without requiring users to restart their task.
- **Perceived Sovereignty:** Assesses whether users trust that their voice data remains under their control. This qualitative metric evaluates confidence that audio is not stored or reused without consent and requires community-centered, ethnographic evaluation methods.
- **Latency:** Evaluates whether system response timing aligns with Guaraní conversational tempo, avoiding both premature interruptions that violate turn-taking norms and prolonged silences that disrupt conversational flow.

4. Discussions and Limitations

The proposed architecture highlights the potential of oral-first language technology, but several challenges extend beyond technical implementation.

4.1. Standardization vs. Lived Reality

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, [n. d.]), daily use frequently involves *jopará* and code-switching (Mortimer, 2006; Estigarribia, 2015; Kellert and Tyagi, 2025). As a result, standardized forms may diverge from lived usage. 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).

4.2. The Data Bottleneck is Specifically Conversational

Beyond general data scarcity, oral-first systems lack conversational audio capturing turn-taking, repair, and shared context. Many endangered languages remain primarily oral and underrepresented in digital resources (Turin, 2012). For Guaraní, existing corpora and scraping-based methods mainly support text-based evaluation rather than spontaneous multi-turn speech (Chiruzzo et al., 2022; Góngora et al., 2021). Speech recognition results from the AmericasNLP shared task further confirm that performance on Guaraní lags significantly behind higher-resource languages (Ebrahimi et al., 2023), while recent ASR work fine-tuning Whisper for Guaraní demonstrates both progress and the continued need for domain-appropriate conversational data (Acevedo Zarza et al., 2024). Future efforts should prioritize community-led collection of conversational data (JournalismAI, 2025).

4.3. Governance and Perceived Control

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 et al., 2024).

5. 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. This framing moves beyond universal, text-first assumptions toward interaction models that reflect human communicative coordination and sociolinguistic reality. 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.

Santiago Rubén Acevedo Zarza, Mateo Andrés Fidabel Gill, Christian Daniel von Lücken Martínez, and Diego Pedro Pinto Roa. 2024. Desarrollo de un sistema de reconocimiento del habla en guaraní: Evaluación de variantes del modelo Whisper y técnicas de mejora de datos. *JAIIO, Jornadas Argentinas de Informática* 10, 1 (2024), 158–166. doi:10.1145/3641234

Essam Alghamdi, Martin Halvey, and Emma Nicol. 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* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 28, 13 pages. doi:10.1145/3640794.3665558

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4218–4222. https://aclanthology.org/2020.lrec-1.520/

Evan Becker. 2024. Multi-Agent Large Language Models for Conversational Task-Solving. arXiv preprint arXiv:2410.22932v1. https://arxiv.org/abs/2410.22932

Stephanie Russo Carroll et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal* (2020).

Siyuan Chen et al. 2025. Multi-Party Conversational Agents: A Survey. arXiv preprint arXiv:2505.18845v1. https://arxiv.org/abs/2505.18845

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache,

Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2098–2107. https://aclanthology.org/2022.lrec-1.226/

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*. American Psychological Association. https://www.cs.cmu.edu/~illah/CLASSDOCS/Clark91.pdf

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G. Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Ivan Meza-Ruiz, John E. Ortega, Alexis Palmer, Rodolfo Zevallos, Kristine Stenzel, Thang Vu, and Katharina Kann. 2023. Findings of the Second AmericasNLP Competition on Speech-to-Text Translation. In *Proceedings of the NeurIPS 2022 Competitions Track (Proceedings of Machine Learning Research, Vol. 220)*. PMLR, 217–232. https://proceedings.mlr.press/v220/ebrahimi23a.html

Bruno Estigarribia. 2015. Jopará and Guaraní in Paraguay (discussion of contact and mixed speech). Cited in Guaraní NLP work as evidence for contact-driven variation..

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guaraní*. UCL Press. https://uclpress.co.uk/book/a-grammar-of-paraguayan-guarani/

Charles A. Ferguson. 1959. Diglossia. *Word* 15, 2 (1959), 325–340. doi:10.1080/00437956.1959.11659702

Joshua A. Fishman. 1967. Bilingualism with and without Diglossia; Diglossia with and without Bilingualism. *Journal of Social Issues* 23, 2 (1967), 29–38. doi:10.1111/j.1540-4560.1967.tb00573.x

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani Corpus of News and Social Media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann (Eds.). Association for Computational Linguistics, Online, 153–158. doi:10.18653/v1/2021.americasnlp-1.16

Ben Hutchinson. 2025. A partnership with The University of Western Australia to improve speech technology for Aboriginal and Torres Strait Islander people's voices. Google Australia Blog. https://blog.google/intl/en-au/company-news/technology/a-partnership-to-improve-speech-technology-for-first-nations-voices/

Instituto Nacional de Estadística (INE), Paraguay. 2024. Día Internacional de la Lengua Materna: Diversidad lingüística en Paraguay. https://www.ine.gov.py/noticias/2298/dia-internacional-de-la-lengua-materna-diversidad-linguistica-en-paraguay Household language-use reporting based on EPHC 2023. Accessed 2026-02-02.

Instituto Nacional de Estadística (INE), Paraguay. 2025. 8 de cada 10 personas utiliza internet en Paraguay (EPH 2017–2024). Reports 81.6% internet use among population aged 10+ in 2024, up from 61.1% in 2017; includes noted exclusions. Accessed 2026-02-02.

Hiroshi Ito. 2012. With Spanish, Guaraní lives: a sociolinguistic analysis of bilingual education in Paraguay. *Multilingual Education* 2, 1 (2012), 6. doi:10.1186/2191-5059-2-6

JournalismAI. 2025. Guarani AI: When building language tech means building community. https://www.journalismai.info/blog/5fcm6ayykhqq7564kbvt9nw92wwmy9 Documents community "mingas" and Guaraní audio dataset efforts. Accessed 2026-02-02.

Olga Kellert and Nemika Tyagi. 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.

Katherine Mortimer. 2006. Guaraní Académico or Jopará? Educator Perspectives and Ideologies on Language in Paraguay. Often cited for discussions of standardization, literacy, and ideologies around Guaraní/Spanish mixing..

Organization of American States (OAS). 1992. Paraguay's Constitution of 1992 with Amendments through 2011. PDF. https://www.oas.org/ext/Portals/33/Files/Member-States/Parag_intro_textfun_eng_1.pdf English translation; consolidated text with amendments through 2011.

Secretaría de Políticas Lingüísticas. [n. d.]. 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. https://spl.gov.py/files/legal/Ley%204251%20-%20bilingue.pdf

Jahanzeb Sherwani, Nosheen Ali, Carolyn Penstein Rosé, and Roni Rosenfeld. 2009. Orality-Grounded HCID: Understanding the Oral User. *Information Technologies & International Development* 5, 4 (2009), 37–49. https://itidjournal.org/index.php/itid/article/download/422/422-1096-2-PB.pdf

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language* 67 (2021), 101178. doi:10.1016/j.csl.2020.101178

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, G. Hoymann, Federico Rossano, Jan P. de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592. doi:10.1073/pnas.0903616106

The University of Western Australia. 2025. First Nations people to benefit from inclusive technology partnership. UWA News. https://www.uwa.edu.au/news/article/2025/february/first-nations-people-to-benefit-from-inclusive-technology-partnership

Mark Turin. 2012. Voices of vanishing worlds: Endangered languages, orality, and cognition. *Análise Social* 205, 47 (2012). https://www.researchgate.net/publication/262778986_Voices_of_vanishing_worlds_Endangered_languages_orality_and_cognition