RESEARCH ARTICLE

SAMRIDDHI Volume 15, Issue 4, 2023 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Voice Interfaces and Vernacular Languages: Bridging the Digital Divide through Conversational AI in Multilingual Societies

Dyuti Banerjee

Department of CSE Koneru Lakshmaiah Education Foundation Green Fields, Guntur , Andhra Pradesh, India

Abstract

Text-centric computing has quietly imposed a language tax on the world's multilingual societies: to participate digitally, users must operate in a dominant script and register that may be neither their spoken language nor their preferred one. Voice interfaces promise to lift this tax by making speech — the one channel nearly universal across literacy levels — the primary mode of interaction. Yet the promise is unevenly kept. Speech technologies perform best for high-resource, standardised languages and degrade sharply for vernaculars, dialect continua, and code-switched speech, threatening to reproduce the very divide they are meant to bridge. This paper analyses voice interfaces as digital-divide infrastructure in multilingual societies. We develop a layered model of the vernacular voice stack — speech recognition, language understanding, dialogue management, content, and speech output — and identify at each layer the specific barriers facing vernacular deployment, from data scarcity and orthographic instability to register mismatch and evaluation blind spots. We consolidate these into a barrier–intervention framework linking technical, design, and governance responses, and discuss deployment lessons from voice-first services in agriculture, health, and government assistance. We argue that inclusive conversational AI is not a smaller version of English conversational AI but a differently shaped problem, requiring community-participatory data practices, dialect-tolerant design, and public investment in language resources as digital public goods.

Keywords: voice user interfaces; conversational AI; vernacular languages; digital divide; low-resource languages; speech recognition; multilingual societies; inclusive design.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2023); DOI: 10.18090/samriddhi.v15i04.13

Introduction

For most of computing history, access has been mediated by text. Keyboards, menus, forms, and search boxes all presume that the user can read and write, and — more restrictively — that they can do so in one of the small set of languages and scripts that platforms support well. In multilingual societies such as India, Nigeria, Indonesia, and the Philippines, this presumption excludes at multiple levels simultaneously: adults with limited literacy in any language; literate speakers whose language has weak digital support; and the very large population whose everyday speech is a vernacular, dialect, or code-switched blend that differs substantially from the standardised register that technology recognises.

Voice interfaces appear to dissolve the problem at its root. Speaking requires no literacy, no keyboard proficiency, and no navigation of visual hierarchies; interactive voice response systems on ordinary phones already reach populations that smartphone apps do not. The rapid progress of neural speech recognition and large language models has strengthened the expectation that conversational AI will become the great equaliser of digital access. This paper subjects that expectation to scrutiny. Our central claim is that voice is necessary but not sufficient: unless the full stack of speech technology is deliberately adapted to vernacular linguistic reality, voice interfaces will serve the standardised registers of dominant languages and leave the divide intact — now hidden behind an interface that appears universal. We contribute (i) a layered analysis of where vernacular

Corresponding Author: Dyuti Banerjee, affiliation, e-mail: dbanerjee@kluniversity.in

How to cite this article: Banerjee, D. (2023). Voice Interfaces and Vernacular Languages: Bridging the Digital Divide through Conversational AI in Multilingual Societies. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 15(4), 74-78.

Source of support: Nil

Conflict of interest: None

deployment breaks down, (ii) a barrier–intervention framework (Table 1) connecting technical, design, and governance responses, and (iii) deployment principles drawn from voice-first services in development contexts.

Background

The Digital Divide as a Language Divide

Digital divide research has moved through three levels: access to devices and connectivity, skills and usage capability, and the outcomes users derive from technology. Language cuts across all three but is most acute at the second and third levels: a user may own a phone and pay for data yet be unable to convert them into benefit because services do not operate in a language they command. Estimates of global linguistic diversity count roughly 7,000 living languages, while mainstream speech products support at most a few dozen with high quality. The gap is not merely one of language count: within a single "supported" language, the standardised variety used in training data can differ from spoken vernaculars in phonology, lexicon, and syntax to the point of mutual difficulty. For speech systems, the unit that matters is not the language but the variety.

Why Vernaculars Are Hard for Speech Technology

Four structural features distinguish vernacular speech from the material on which speech systems are typically trained. First, data scarcity: transcribed audio in vernacular varieties is limited, and what exists often reflects broadcast or read speech rather than conversation. Second, orthographic instability: many vernaculars lack a standard written form, or are written in multiple scripts, which complicates transcription, model training, and evaluation alike. Third, code-switching: multilingual speakers routinely mix languages within a sentence, while most recognition pipelines assume a single active language. Fourth, dialect continua: varieties shade into one another geographically, so a system trained on one point of the continuum degrades gracefully in engineering terms but unpredictably in user-experience terms. These features interact: scarce data plus unstable orthography makes benchmark construction itself contested, which in turn hides performance gaps from developers.

Voice-First Services in Development Practice

A substantial ICT4D literature documents voice services built for low-literacy and low-connectivity contexts: community radio-style voice forums for farmers, IVR health hotlines, and voice-based job and information marketplaces. Three durable lessons emerge. Human-recorded prompts in the local dialect earn trust that synthetic voices historically struggled to match. Constrained, task-focused dialogues succeed where open-ended assistants flounder, because narrow domains reduce the recognition and understanding burden. And mediation persists: even with voice, family members and community intermediaries assist first-time users, so systems should support rather than resist assisted use. These lessons predate modern conversational AI but remain design-relevant, because they concern trust and task structure, not model capability.

The Vernacular Voice Stack

We model a voice service as five layers, each of which can independently exclude vernacular speakers. Automatic speech recognition (ASR) converts speech to a machine representation; its errors for under-represented varieties are systematic, concentrating on precisely the users the service is meant to include. Natural language understanding (NLU) maps utterances to intents; models trained on standard registers misread vernacular idiom, indirect speech acts, and politeness forms. Dialogue management structures the interaction; turn-taking conventions, barge-in tolerance, and repair strategies that suit fluent technology users can confuse first-time users, whose responses to open prompts are often narrative rather than slot-shaped. The content layer determines whether anything worth hearing exists in the language at all — recognition without vernacular content is an empty pipe. Finally, speech output (TTS) must render responses in a voice and register the community accepts; a correct answer in an alien accent or formal register can fail pragmatically. Table 1 organises the principal barriers at each layer and matches them to interventions across three response classes.

Analysis: Three Deployment Principles

Treat Language Resources as Digital Public Goods

The economics of vernacular speech technology are unfavourable to purely commercial provision: data collection costs are high, per-language markets are small, and benefits accrue broadly. This is the classic profile of a public good. Governments and multilateral funders can correct the under-supply by financing open speech corpora, pronunciation lexicons, and evaluation sets for their countries' varieties, with licences that permit both public and commercial reuse. National language-resource missions — building open datasets and baseline models across constitutional and community languages — represent this approach, and their central design questions are governance ones: who consents to and is compensated for contributed speech, who decides which varieties are prioritised, and how communities retain a say in downstream use.

Table 1: Barrier–intervention framework for vernacular voice interfaces

Stack Layer	*Principal Barriers for Vernaculars*	*Technical Interventions*	*Design & Governance Interventions*
Speech recognition (ASR)	Scarce transcribed audio; code-switching; dialect variation; noisy, shared-device acoustic conditions	Transfer learning from related languages; self-supervised pre-training on untranscribed audio; code-switch-aware modelling	Community-participatory data collection with consent and compensation; dialect-stratified error reporting
Language understanding (NLU)	Register mismatch; idiom and indirect speech acts; unstable orthography in training text	Intent models trained on spoken-form transcripts; normalisation layers tolerant of spelling variance	Local-language annotation teams; escalation paths when confidence is low
Dialogue management	Narrative (non-slot) answers; unfamiliar turn-taking; abandonment after recognition failure	Mixed-initiative dialogues; explicit confirmation with read-back; graceful repair scripts	Task-focused scope; human handoff as a designed state, not a failure state
Content	Little service content authored in the vernacular; translation that imports formal register	Retrieval over locally authored corpora; controlled-language authoring tools	Fund vernacular content creation as a public good; community editorial oversight
Speech output (TTS)	No natural-sounding voices for the variety; register and accent mistrust	Low-resource voice building from small studio corpora; recorded human prompts for critical messages	Voice-talent selection with community input; disclosure that the voice is synthetic
Cross-cutting: evaluation	Benchmarks in standard varieties hide vernacular failure; aggregate metrics mask group-level gaps	Variety-disaggregated test sets; in-the-wild task-success measurement	Procurement and funding conditioned on disaggregated reporting

Design for the Variety, Not the Language

A service that claims to support a language but recognises only its broadcast standard makes an implicit promise it cannot keep, and the users it fails will often blame themselves. Deployment targets should therefore be specified at the level of varieties: which dialect regions, which code-switching patterns, which registers. This has cascading consequences — test sets must be stratified by variety; error budgets must be set per stratum; and dialogue design must include repair strategies that do not require the user to shift into the standard register (for example, offering recognition alternatives aloud rather than asking the user to "repeat clearly"). Where variety coverage is genuinely infeasible, honesty is a design feature: constraining the domain and stating clearly what the service can do preserves trust better than a broad assistant that fails unpredictably.

Keep Humans in the Loop by Design

In high-stakes domains — health advice, government entitlements, financial transactions — vernacular voice systems should be built as triage-and-handoff architectures: the conversational layer resolves routine queries and routes the remainder, with full context, to human operators. This is not a temporary concession to imperfect technology but a standing architectural principle, for two reasons. First, recognition and understanding errors for vernacular speakers are systematic, so purely automated paths institutionalise unequal service quality. Second, the trust research is consistent: knowing that a human is reachable increases willingness to attempt the automated path at all. The handoff must preserve state — users who repeat their story from the beginning after an automated failure rarely return.

Discussion

Two tensions complicate the optimistic reading of recent progress. The first is the aggregation illusion. Multilingual foundation models report impressive average performance across many languages, but averages are exactly the wrong statistic for divide analysis: what matters is the tail — performance for the varieties spoken by the least-served users. Without variety-disaggregated evaluation, improvement at the mean can coexist with stagnation at the tail, and public claims of language coverage can outrun lived experience. The second is the extraction risk. Community speech data has become valuable, and collection efforts can slide into extractive patterns in which communities contribute voice and receive neither capability nor compensation. Participatory frameworks — consent that is specific and revocable, benefit-sharing, community governance of corpora — are the governance counterpart of the public-goods argument in Section 4.1. There is also a cultural dimension that engineering framing tends to miss: for many communities, language technology is entangled with language vitality. A well-built vernacular voice interface can raise a variety's prestige and everyday utility, while its absence signals — to young speakers especially — that the language has no digital future. Voice interfaces are therefore instruments of language policy whether or not their builders intend it.

Conclusion

Voice interaction is the most plausible route to digital inclusion for popultions excluded by text, but the route is not automatic. This paper has argued that inclusion

depends on the entire vernacular voice stack — recognition, understanding, dialogue, content, and output — and that each layer harbours barriers that default, standard-register development will not remove. The barrier–intervention framework consolidates technical responses (transfer learning, code-switch modelling, low-resource voice building) with the design and governance measures (participatory data practices, variety-level targets, disaggregated evaluation, human handoff) without which the technical responses under-deliver. Future work should validate the framework against live deployments across domains, develop shared variety-stratified benchmarks for major dialect continua, and examine the political economy of language-resource missions comparatively across countries. The test of conversational AI in multilingual societies is not whether it can converse, but with whom: a divide-bridging technology must be measured at the margins of the speech community, where the divide actually lives.

References

[1] Cherladine, K., Rallabandi, B. R., Goel, A., Kaushik, K., Dhillon, A., & Soni, M. (2025). Generative adversarial defense models for simulated cyberattack scenarios in virtual networks. In 2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICDISS68238.2025.11320686

[2] Rallabandi, B. R. (2020). MEC-native 5G systems: Orchestration algorithms for ultra-low latency cloud-edge integration. International Journal of Intelligent Systems and Applications in Engineering, 8(4), 398–408. https://ijisae.org/index.php/IJISAE/article/view/7889

[3] Rallabandi, B. R. (2018). Joint deployment and operational energy optimization in heterogeneous cellular networks under traffic variability. International Journal of Communication Networks and Information Security, 10(3), 638–647. https://ijcnis.org/index.php/ijcnis/article/view/8604

[4] C. H. Neetha and M. P. Kantipudi, "Adaptive Edge-Based Bilinear Interpolation for Smart Healthcare," in IET Conference Proceedings CP824, vol. 2022, no. 26, Stevenage, U.K.: The Institution of Engineering and Technology (IET), 2022, pp. 7–13.

[5] R. Aluvalu, R. Venkatachalam, V. Uma Maheswari, R. J. Anandhi, M. V. V. Kantipudi, and S. Prashanth Mallellu, "IWHO-Based Cluster Head Selection for Vehicle-to-Vehicle Communication in Intelligent Transport System," International Journal of Transport Development and Integration, vol. 8, no. 2, pp. 154–166, 2024

[6] S. Velamuri, S. K. Sudabattula, M. V. V. Prasad Kantipudi, and N. Prabaharan, "Q-Learning Based Commercial Electric Vehicles Scheduling in a Renewable Energy Dominant Distribution Systems," Electric Power Components and Systems, pp. 1–14, 2023

[7] Singh, M., Rallabandi, B., Gattupalli, K. K., & Sai Medarametla, M. (2025). Autonomous private 5G field area networks: Achieving secure, scalable, and self-healing smart grid communications. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448712

[8] Desai, A. B., Rallabandi, B., Gattupalli, K. K., & Medarametla, M. S. (2025). Agentic AI for rural connectivity and spectrum-aware networks to power smart agriculture ecosystems. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448879

[9] Goyal, S., Rallabandi, B., Gattupalli, K. K., & Medarametla, M. S. (2025). Digital twin-enabled context-aware networks: Enhancing smart grid resilience through predictive intelligence. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448880

[10] Tripathi, N., Gattupalli, K. K., Rallabandi, B., & Medarametla, M. S. (2025). AI-native Cloud-RAN orchestration for enterprise private 5G using digital twin models. In 2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON) (pp. 851–857). IEEE. https://doi.org/10.1109/UPWIECON67212.2025.11390348

[11] Goyal, S., Gattupalli, K. K., Rallabandi, B., & Medarametla, M. S. (2025). Integrating terrestrial and non-terrestrial networks for reliable rural coverage in agriculture and smart grids. In 2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON) (pp. 846–850). IEEE. https://doi.org/10.1109/UPWIECON67212.2025.11390331

[12] N. Pachauri, V. Suresh, M. V. V. Prasad Kantipudi, R. Alkanhel, and H. A. Abdallah, "Multi-Drug Scheduling for Chemotherapy Using Fractional Order Internal Model Controller," Mathematics, vol. 11, no. 8, Art. no. 1779, 2023,

[13] M. V. V. Prasad Kantipudi, S. Vemuri, N. S. Pradeep Kumar, S. Sreenath Kashyap, and S. Eslamian, "Proposing Model for Water Quality Analysis Based on Hyperspectral Remote Sensor Data," in Handbook of Hydroinformatics, Elsevier, 2023, pp. 317–324.

[14] Rana, M., Srinivas, S., Jamili, L. K., Jaiswal, I. A., Nakka, S., & Kasetti, S. (2025, May). Real-Time Monitoring and Prediction of Blood Sugar Levels in Diabetic Patients with Functional Models. In 2025 International Conference on Engineering, Technology & Management (ICETM) (pp. 1-6). IEEE.

[15] Jaiswal, I. A. (2023). Intelligent Cybersecurity Framework for Large-Scale RESTful Service Architectures. International Journal of Research Radicals in Multidisciplinary Fields, ISSN: 2960-043X, 2 (1), 178–184. Retrieved from https://www. researchradicals. com/index. php/rr/article/view, 252.

[16] Jaiswal, I. A. (2023). Multilingual and culturally adaptive AI models for global education platforms. International Journal for Research in Education (IJRE), 12(9), 17–27. https://doi.org/10.63345/ijre.v12.i9.1

[17] M. S. Prashanth, R. Aluvalu, and M. V. V. Kantipudi, "Enhancing Health Product Traceability on the Blockchain: A Novel Approach for Supply Chain Management Inspection to AI," EAI Endorsed Transactions on Pervasive Health & Technology, vol. 10, no. 1, 2024.

[18] H. K. Jani, M. V. V. Prasad Kantipudi, G. Nagababu, D. Prajapati, and S. S. Kachhwaha, "Simultaneity of Wind and Solar Energy: A Spatio-Temporal Analysis to Delineate the Plausible Regions to Harness," Sustainable Energy Technologies and Assessments, vol. 53, Art. no. 102665, 2022

[19] Jaiswal, I. A. (2022). Natural language processing for security policy and log analysis. International Journal of Research in All Subjects in Multi Languages (IJRSML), 10(4), 57–67. https://doi.org/10.63345/ijrsml.v10.i4.1

[20] Khan, S. (2019). Sales force security compliance: An in-depth

study of GDPR, HIPAA, and PCI-DSS enforcement in cloud-based CRM systems and their implications for global enterprises. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(9), 2211–2219. https://doi.org/10.15662/IJAREEIE.2019.0809010

[21] S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.

[22] S. Rani, D. Ghai, and S. Kumar, "Reconstruction of wire frame model of complex images using syntactic pattern recognition," in IET Conference Proceedings CP791, vol. 2021, no. 11, pp. 8–13, 2021.

[23] Swathi, S. Kumar, S. Rani, A. Jain, and R. K. M. V. N. M., "Emotion classification using feature extraction of facial expression," in Proceedings of the 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 283–288, 2022.

[24] S. Choudhary, S. Kumar, S. Gowroju, M. Gulhane, and R. Sri Lakshmi, Eds., Genomics at the Nexus of AI, Computer Vision, and Machine Learning. Hoboken, NJ, USA: John Wiley & Sons, 2024.

[25] S. Kumar, S. Choudhary, S. Gowroju, and A. Bhola, "Convolutional neural network approach for multimodal biometric recognition system for banking sector on fusion of face and finger," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 251–267, 2023.

[26] S. Choudhary, S. Kumar, M. Gulhane, and M. Kumar, "Secured automated certificate creation based on multimodal biometric verification," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 269–281, 2023.

[27] Jaiswal, I. A. (2024). AI-powered observability and incident prediction in distributed enterprise platforms. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(1), 1–14. https://doi.org/10.63345/sjaibt.v1.i1.201

[28] S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.

[29] Jaiswal, I. A. (2023). High-Performance AI-Augmented Content Management Systems for Distributed Clouds. International Journal of Multidisciplinary Innovation and Research Methodology, ISSN: 2960-2068, 2 (2), 90–97. Retrieved from https://ijmirm. com/index. php/ijmirm/article/view, 243.

[30] Khan, S. (2018). The role of zero-trust models in database security: Eliminating implicit trust and enforcing continuous verification in enterprise data access systems. International Journal of Research in Electronics and Computer Engineering, 6(1), 1622–1628.

[31] Jaiswal, I. A. (2024). Self-healing REST services using artificial intelligence in multi-cloud environments. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(3), 1–7. https://doi.org/10.63345/sjaibt.v1.i3.201

[32] V. Saini, A. Jain, Anurag, and M. V. V. Prasad Kantipudi, "An Efficient Approach for Improving the Performance of Autonomous Vehicle Using Advanced Computer Vision," in International Conference on Soft Computing and Pattern Recognition, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 164–174.

[33] Khan, S. (2017). The role of privacy-preserving techniques in database security: Secure multi-party computation, differential privacy, and homomorphic encryption. The Research Journal, 3(4), 29–35.

[34] Khan, S. (2016). A study of data provenance and integrity in database security: Ensuring authenticity, non-repudiation, and accountability in data lifecycle management. International Journal of Research in Electronics and Computer Engineering, 4(3), 180–186.