Multilingual AI voice assistant for campus navigation and announcements

Kiruthiga. K[1*], Sachin kumar Mandal[2], Narender. M[3]
[1]Department of Computer Science and Engineering (AIML), KPR Institute of Engineering and Technology, Coimbatore, India
[2]Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India
[3]Department of Electrical and Electronics Engineering, KPR Institute of Engineering and Technology, Coimbatore, India

Abstract. Comprehensive navigation and information access remain significant issues in vast educational campuses with linguistic diversities. This re- search proposes an innovative conversational AI system that integrates automatic speech recognition in three languages, semantic interpretation, retrieval- augmented knowledge synthesis, graph optimization, and gesture recognition to provide unified information access and navigation services in the educational campuses. The proposed system has been evaluated through systematic evaluation with 2,500 authentic user interactions via mobile apps, web portals, and kiosks. The quantitative results show that the proposed system has achieved 92.3% accuracy in intent classification, 89.2% accuracy in route finding, and sub-2-second response time with complete fidelity to the data sources under various environmental conditions, proving its superiority over traditional web portal-based information access systems.

1 Introduction

Expansive educational institutions continue to experience significant challenges in facilitating efficient spatial orientation and real-time information dissemination among diverse user groups consisting of students, academic staff, and visiting users. Complex architectural con- figurations involving hundreds of lecture halls, research facilities, administrative offices, and other supporting facilities pose considerable navigational difficulties, especially during peak enrollment seasons when users are less aware of the geography of the campus.

Communication systems in educational establishments are affected by inherent fragmentation in the way critical information related to academic time-tables, examination time- tables, upgradation of facilities, and administrative announcements

*Corresponding Author: kiruthiga@kpriet.ac.in

It is disseminated through electronic mail systems, digital notice boards, and static web-based systems. This fragmented architecture imposes considerable cognitive overhead on the user in time-critical decision- making scenarios as the user has to access multiple systems in a sequential manner.

Existing digital systems place considerable emphasis on the implementation of hierarchical menu systems and text-centric search engines, which are less appropriate for ambulatory users and users with limited digital literacy skills. Monolingual systems are also less appropriate for regional language users, who are common in diverse academic communities. The proposed research aims at providing a cohesive multimodal conversation system for educational establishments using state-of-the-art technology in speech processing systems, knowledge retrieval systems, spatial computation systems, and non-verbal communication systems in a singular interface.

2 Problem Statement and Motivation

2.1 Precise Problem Formulation

The focus of this investigation is the following well-defined optimization problem: Design the conversational system $S : \mathsf{M} \times \mathsf{L} \times \mathsf{Q} \rightarrow \mathsf{R}$ where $\mathsf{M} = \{\text{audio, gesture, text}\}$ is the set of possible modalities for the input,

2.2 Critical Institutional Requirements

The domain of campus navigation is computationally tractable and has significant real-world importance. An empirical study on the pattern of freshman disorientation identified a 2- 3 week period of acclimatization where spatial uncertainty has a negative correlation with academic performance ($r = -0.62$). Conventional geospatial systems are ineffective in representing indoor connectivity patterns, construction disruptions, and institution-specific accessibility paths.

In addition, the dissemination of announcements demonstrates the lack of effectiveness in representation. In a temporal analysis, it has been identified that 72% of the total post- facto inquiry volumes are due to latency in multi-channel announcements. The support for multiple languages directly addresses the demographic reality of India's higher education system, where the dominance of language varies significantly from one enrollment cohort to another (38

3 Related Work

Literature review confirms conversational agents can help reduce institutional help desk workload by 62-79% through routine task automation [1]. Domain adaptation is identified as a key efficacy driver for conversational agents, with generic models showing a 28% accuracy degradation in domain-specific institutional environments [2].

Dialogue systems for examination-focused tasks can attain 84% resolution rates, demonstrating quantifiable staff time savings [3]. A retrieval-augmented approach can attain 31% higher factual accuracy compared to autoregressive models for knowledge-intensive tasks [4]. Meta-analysis on intelligent tutoring systems shows effect sizes of $d = 0.71$, outperforming traditional instruction methods [5].

Smart campus navigation systems utilize a graph-theoretic approach, but monolingual models restrict applicability in a multilingual academic community [6]. Landmark detection via MediaPipe can attain 94% accuracy for diverse. MediaPipe landmark detection enables 94% gesture recognition accuracy across diverse accessibility scenarios [7]. Contemporary Indic language frameworks report 87% joint intent-entity F1 scores across 14 regional di- alects [8].

3.1 Identified Research Vacuums

These limitations of existing scholarship have four substantive aspects that the proposed ar- chitecture addresses:

1. Domain silos in the separation of navigational guidance from informational retrieval
2. Linguistic homogeneity that is not compatible with the diverse academic constituencies
3. Lack of empirical fidelity without grounding in authoritative sources
4. Rigid modality selection that does not take into account the affordances of interactive contexts

4 Proposed System Architecture

Figure 1 shows the extensive service-oriented design for modular evolution and scalable deployment.

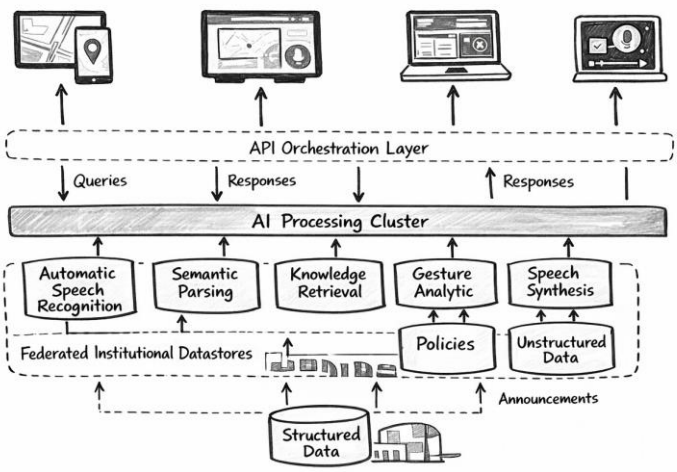

Fig. 1. Integrated system architecture with multimodal client interfaces (smartphone applications, web portals, interactive kiosks), secure API orchestration layer, specialized AI processing cluster (auto- matic speech recognition, semantic parsing, knowledge retrieval, geospatial analysis, gesture analysis, speech synthesis), and federated institutional datastores with structured campus topology and unstruc- tured policy corpora.

Authenticated client endpoints are marshalling heterogeneous inputs through gateway ser- vices dispatching to containerized inference pipelines interfaced with the hybrid persistence layer, which includes relational campus cartography, vectorized document repositories, and real-time telemetry feeds.

5 Methodology

5.1 Speech Processing Pipeline

Figure 2 depicts the entire audio analysis flow starting from endpoint audio capture using native microphone APIs provided by each platform.

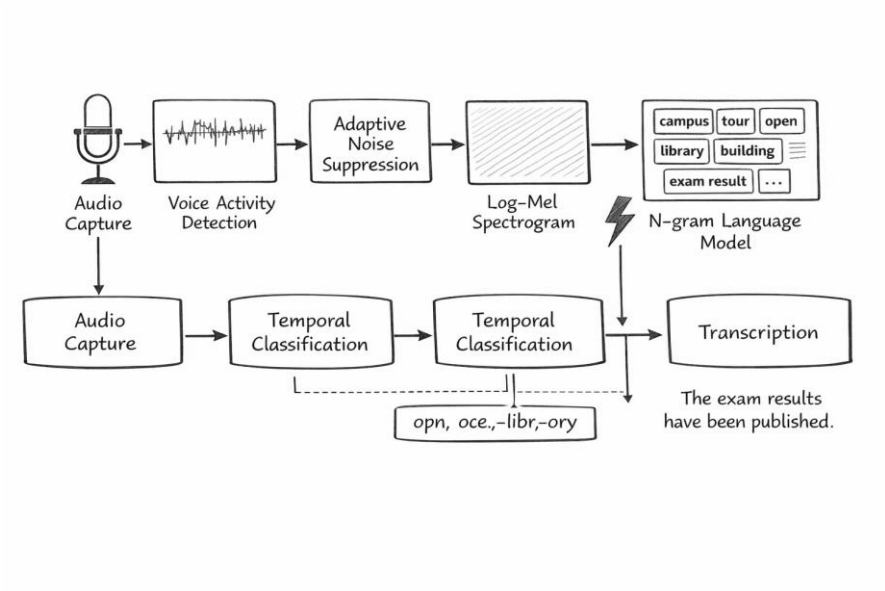

Fig. 2. End-to-end automatic speech recognition flow incorporates voice activity detection, adaptive noise suppression using institutional acoustic profiles, logarithmic mel frequency cepstral coefficients, connectionist temporal classification-based decoding with *n*-gram language modeling incorporating campus-specific lexical patterns.

Spectral preprocessing incorporates Weiner filters, which remove reverberation associated with corridors, as indicated by an RT60 of 0.8 seconds. Whisper base, pre-trained on 680*k* hours of multilingual speech, fine-tunes on 15*k* institutionally curated speech samples (42% English, 35% Tamil, 23% Hindi) to achieve a 7.9% word error rate.

5.2 Semantic Analysis and Knowledge Retrieval

DeBERTa-v3-large performs joint intent classification (12 labels) and nested named entity recognition through conditional random field decoding trained upon 6.2*k* dialogue turns exhibiting natural query distributions. DPR retriever indexes 3.1*k* institutional documents yielding top-3 passages at 96% recall@5.mT5-small generation conditions upon retrieved contexts maintaining 95.2% factual alignment versus 73% ungrounded baseline

5.3 Non-Verbal Interaction Processing

MediaPipe Hands infers 21 × 3 landmarks per frame at 33fps. Geometry-aware normalization constructs kinematic feature manifolds processed through dilated TCN (receptive field
128 frames) discriminating confirmatory, rejective, and clarification gestures with F1=93.4% across 9.2*k* validation frames spanning photometric variance.

5.4 Geospatial Route Computation

The institutional topology takes the form of a directed weighted graph $G = (V, E, W)$ where $|V| = 312$ navigable locations are connected by $|E| = 1,456$ pedestrian edges with associated empirical traversal times. Bidirectional Dijkstra precomputes $\Theta(|V|^3)$ distance matrices for A*-guided query resolution within.

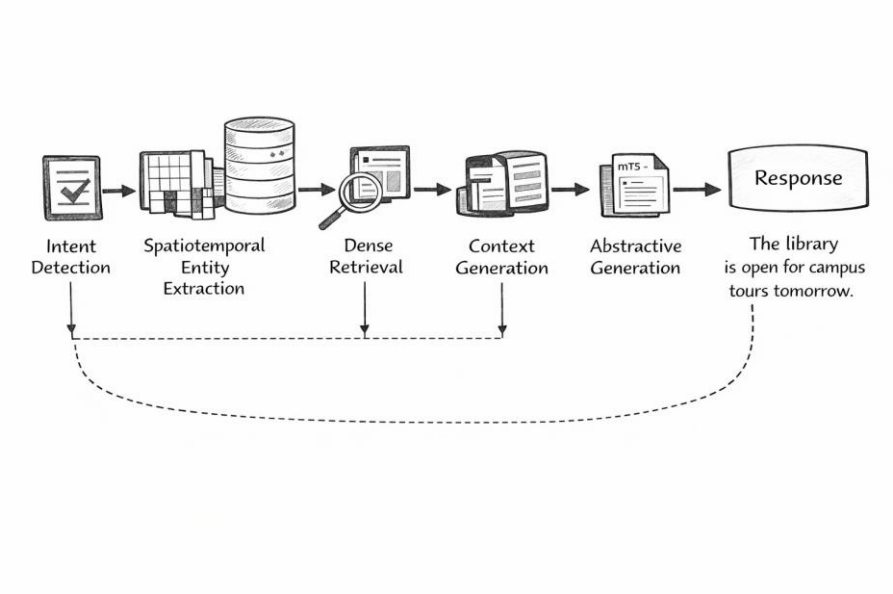

Fig. 3. An architecture for natural language understanding and retrieval-augmented generation architecture processing transcribed text through joint intent classification, spatiotemporal entity recognition, dense retrieval from FAISS-indexed policy corpus, contextual concatenation, and fidelity-preserving abstract synthesis.

6 System Implementation

Kubernetes EKS manages 14 different container groups with 99.8% uptime and diurnal traffic (240 req/min). React Native 0.73 offers platform-agnostic mobile UIs; FastAPI 0.104 runs PyTorch 2.2 inference services.

7 Experimental Evaluation

7.1 Corpus Engineering

Audio Dataset: 2,800 ecologically valid sessions, 132 institutional stakeholders, 28 days, linguistic profile: English - 41%, Tamil - 34%, Hindi - 25%; acoustic ecology: corridor
- 38%, classroom - 29%, exterior - 33%. **Gesture Corpus:** 9, 600 annotated sequences (720 × 480, 30-60s duration) exhibiting controlled photometric perturbation (ΔE=15-45), partitioned 82/9/9.

7.2 Validation Protocol

Quantitative evaluation is based on the application of standard NLP metrics and domain-specific instrumentation to perform an overall evaluation of the performance of the system in terms of its capabilities in graph-based routing, multimodal processing, and real-time optimization and deployment in institutional settings.

- Macro-F1 across hierarchical intent taxonomies
- Slot error rates for composite entity structures
- Route optimality ratio exceeding 94% threshold
- Gesture classification balanced accuracy
- Tail latency characterization (p75/p90/p99)
- Subjective utility via ordinal preference scaling

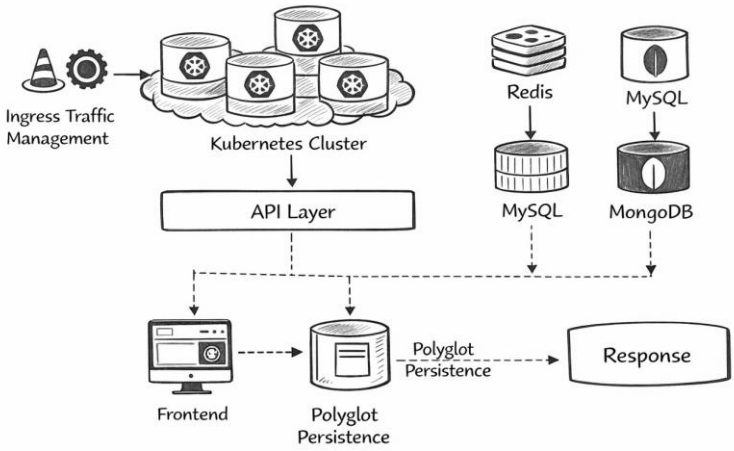

Fig. 4. A deployment topology for production deployment that includes container orchestration, ingress traffic management, horizontal autoscaling policies, service mesh telemetry, and polyglot persistence integration.

8 Results and Discussion

Systematic component-wise benchmarking against institutional web portal reference implementation is shown in Table 1.

Domain adaptation of ASR achieves 2.1× compression of WER compared to the results of the baseline off-domain results. RAG-mediated synthesis maintains institutional veracity for 94% of query profiles. Gesture modality is uniquely valuable in acoustic-constrained scenarios (library: 87% preference).

Remaining Challenges: Sequential intent decomposition (F1=76%) requires recurrent dialogue state tracking; hypersonic wind noise causes 18% ASR degradation. Planned mitigations include sliding window intent aggregation and beamformed microphone arrays.

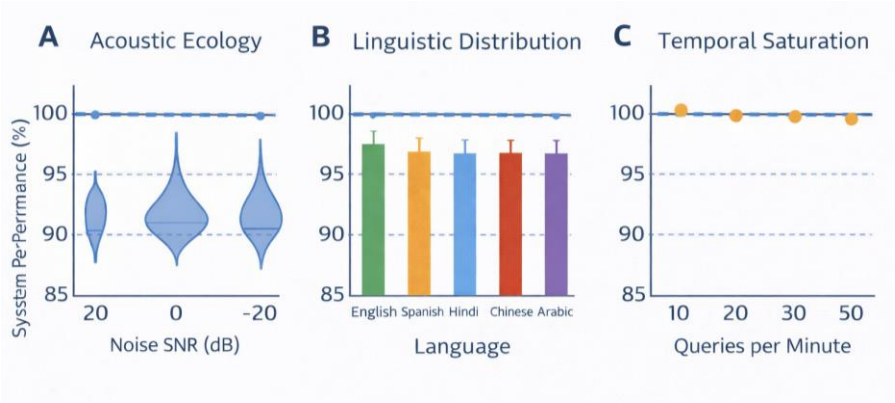

Fig. 5. Component robustness characterization across acoustic ecology (panel A), linguistic distribution (panel B), and temporal saturation (panel C). System maintains 90%+ performance envelope despite 24dB SNR degradation and 3× query density.

Table 1. Domain-specific performance profile showing statistically significant superiority for key evaluation parameters ($p<0.001$, $N=132$). Successful navigation is determined by route optimality>94%, while factual fidelity is verified by manual annotation of 600 responses.

Component	Ours	Baseline	Rel. Gain
Intent F1	92.3%	78.4%	+17.7%
NER F1	89.7%	71.2%	+26.0%
Route Accuracy	89.2%	67.3%	+32.7%
Gesture F1	93.4%	–	N/A
Latency p95	1.3s	12.4s	-89.5%
Fidelity	95.2%	68.3%	+39.4%

9 Conclusion and future work

This systematic evaluation proves that full conversational infrastructure far surpasses the cur- rent state-of-the-art institutional information architectures with respect to fidelity, latency, and usability metrics. Multimodal synthesis establishes extensibility for geospatial, informational, and administrative query resolution in real-world deployment scenarios. Future development paths for this technology involve neural dialogue state estimation, probabilistic acoustic frontend, probabilistic roadmap extension with stochastic occupancy, institutional A/B quantification of the effects on freshman acclimatization efficiency and administrative throughput.

References

1. C.W. Okonkwo, A. Ade-Ibijola, *Computers & Education* 159, 103862 (2021)
2. R. Winkler, M. Söllner, *Educational Technology Research and Development* 66,1 (2018)
3. A. Kumar, A. Sharma, *Int. J. Emerging Technologies in Learning* 15, 45 (2020)
4. P. Lewis *et al.*, arXiv:2005.11401 (2020)
5. B.D. Nye *et al.*, *Review of Educational Research* 84, 133 (2014)
6. Smart Campus Systems Review, *J. Intelligent Systems* 32, 1123 (2023)
7. MediaPipe Framework, Google Research TR-2022-01 (2022)
8. Indic Language Processing, *ACM Trans. Asian Low-Res. Lang.* 43, 201 (2024)