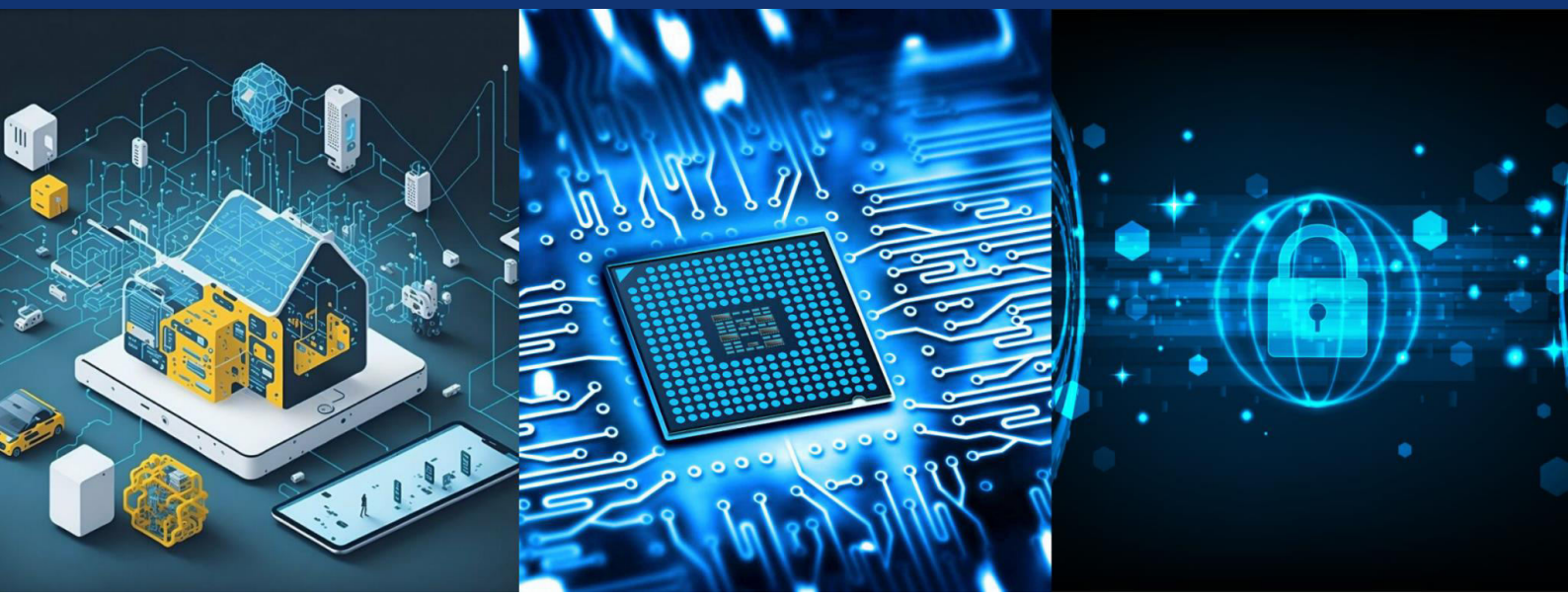
IJIRCCE

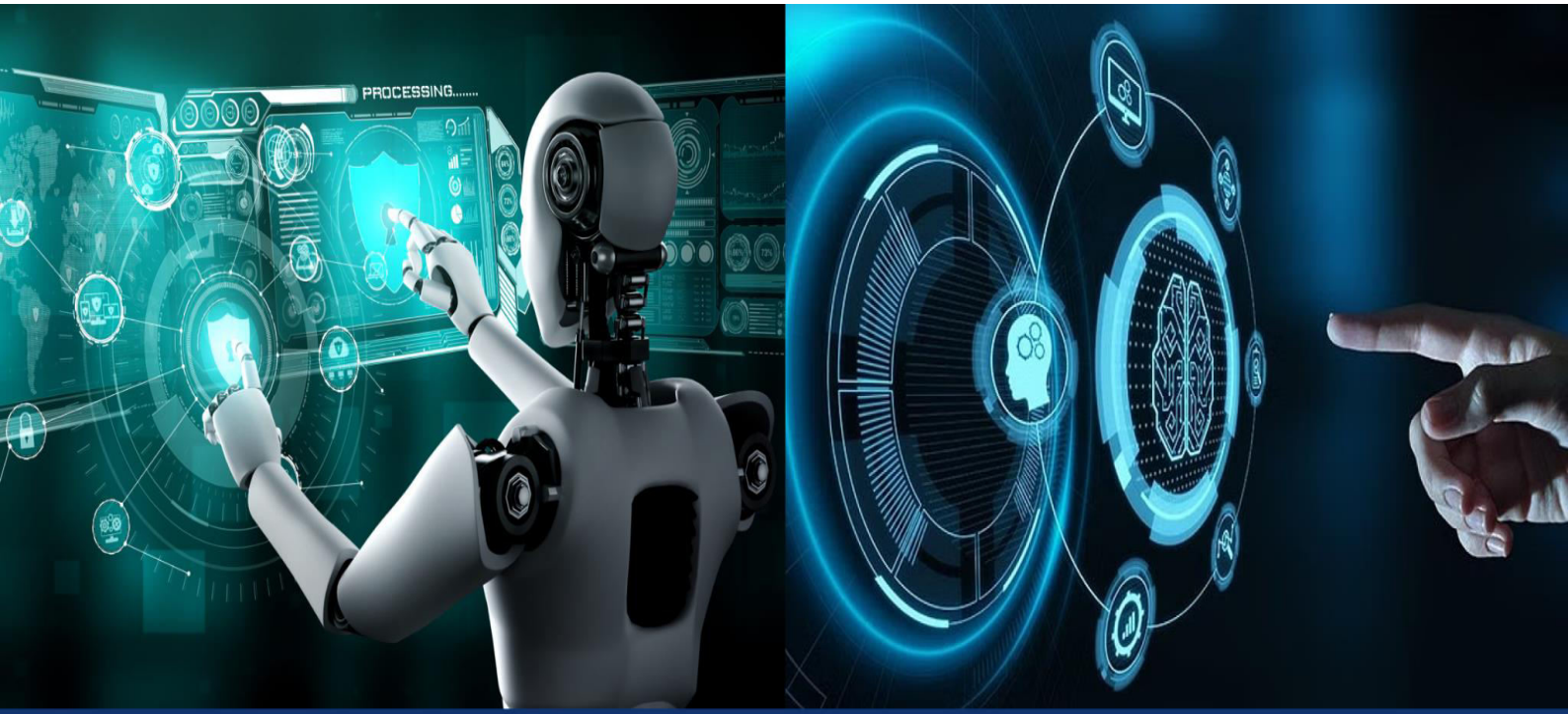
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 14, Issue 4, April 2026

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AI-Enabled Conversational IVR System Using Generative AI for Intelligent Customer Interaction

Manjula P, Santhosh MR[1], Shreya N Gowda[2], Smruti S[3], Suha Noorain HI[4]

Assistant Professor, Dept. of CSE, Jain Institute of Technology, Davangere, Karnataka, India[1]

Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India[2,3,4]

ABSTRACT: A Conversational IVR (CI-IVR) system designed to improve traditional telephony customer service by enabling more natural, personalized automated voice interactions. Key features include integration of large language models, multilingual transformer-based natural language understanding, and neural speech synthesis. The system is cloud-native and modular, with components for noise-robust speech recognition, intent classification, generative response synthesis, sentiment analysis, text-to-speech, and omnichannel dispatch.

In tests involving 5,000 calls and simulated 18,400 call interactions, the system achieved 94.7% intent detection accuracy, 870 ms average latency, and more than doubled call containment compared to traditional DTMF systems. It also succeeded in 88.2% of multilingual queries with code-switching across six Indian languages.

Overall, the CI-IVR framework presents a robust, scalable, and linguistically adaptive solution for next-generation automated contact center services. If you want, I can help summarize, explain specific parts, or assist with related tasks.

KEYWORDS: Conversational AI; Interactive Voice Response (IVR); Large Language Models; Natural Language Processing; Automatic Speech Recognition; Generative AI; Few-Shot Intent Classification; Multimodal Sentiment Analysis; Multilingual NLP; Low-Latency Voice AI

I. INTRODUCTION

The rapid expansion of digital commerce, financial services, and healthcare has driven a growing demand for immediate and personalized customer support. Contact centers act as the main point of interaction between organizations and their customers, traditionally relying on human agents and automated Interactive Voice Response (IVR) systems. Although human agents provide strong contextual understanding, scaling their deployment is cost-prohibitive for enterprises. Conventional IVR systems, which use Dual Tone Multi-Frequency (DTMF) inputs and rigid menu hierarchies, offer some relief but suffer from significant usability challenges. Studies show many callers abandon IVR calls before resolution due to cognitive overload from complex menus, inability to process natural language variations, lack of session memory, and linguistic limitations that exclude non-English speakers. These shortcomings result in revenue loss, higher escalation costs, and diminished brand reputation.

Recent advances in transformer-based language models, deep learning ASR, neural text-to-speech synthesis, and efficient fine-tuning methods like LoRA have opened new possibilities for reimagining IVR systems. However, most current enterprise upgrades address only individual components, maintaining outdated rule-based frameworks. This paper bridges that gap by presenting a comprehensive Conversational IVR (CI-IVR) system that integrates state-of-the-art AI technologies within a unified, production-ready microservices architecture.

The key research contributions include: (i) the design and validation of a modular CI-IVR system combining large language model-based generative response synthesis with telephony infrastructure; (ii) development of a multilingual few-shot intent classification pipeline achieving 94.7% accuracy in realistic telephony conditions; (iii) creation of a real-time multimodal sentiment fusion system for emotion-aware dynamic call routing; (iv) achieving sub-1000 millisecond end-to-end latency through speculative decoding, streaming text-to-speech, and asynchronous processing; and (v) empirical evidence showing a 121.7% increase in call containment and a 104.8% improvement in customer satisfaction (CSAT) compared to traditional DTMF-based IVR systems.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

A review of prior work reveals a diverse yet fragmented research landscape addressing various components of conversational IVR systems. The following synthesis highlights key contributions and gaps that the present work aims to address.

A. Evolutionary Trajectory of Telephony Automation

Kapoor and Mehta [2] identified three phases in telephony automation: DTMF-based menus, NLP-enhanced keyword matching, and LLM-driven generative systems. While their analysis underscores NLP's role in improving first-call resolution, it does not explore architectural demands for generative pipelines operating within strict telephony latency constraints—a challenge directly tackled in this study.

B. Intent Classification in Telecommunications IVR

Ramachandran et al. [3] implemented a transformer-based intent classifier at a telecom operator, reducing agent transfers by 38%. However, their closed-domain taxonomy lacked out-of-domain fallback mechanisms, sentiment awareness, and multilingual support, limiting its applicability in broader, real-world scenarios.

C. Bidirectional Recurrent Architectures for Spoken Language Understanding

Huang and Chen [4] utilized a Bidirectional LSTM model achieving 87.3% intent accuracy on multi-domain call center data. Despite strong results, the approach demands large annotated datasets, posing practical challenges in dynamic enterprise environments where intent categories frequently evolve and annotation is costly.

D. End-to-End AI-IVR Deployment in Retail Banking

Venkataraman and Pillai [5] described an AI-IVR system using Dialogflow for intent recognition combined with deterministic responses. Its English-only design restricts effectiveness in linguistically diverse markets like India, where code-switching between English and regional languages is common.

E. Prototypical Networks for Low-Resource Intent Classification

Liu et al. [6] proposed a prototypical network meta-learning framework for few-shot intent classification with promising results on clean text. However, its robustness under telephony-grade acoustic distortions, including codec noise and accent variability, remains untested.

F. Domain-Specific IVR in Clinical Environments

Sharma and Desai [7] developed an intelligent IVR for hospital appointment scheduling and triage. The system's rigid slot-filling dialogue limited its ability to handle natural multi-symptom descriptions typical of patient interactions

.

G. Theoretical Foundations and Supporting Research

Jurafsky and Martin [8] provide essential insights into transformer architectures, coreference resolution, and dialogue state tracking critical for coherent multi-turn dialogue. Vinyals and Le [9] demonstrated encoder-decoder models for open-domain response generation, though latency and hallucination risks persist. Yu and Deng [10] quantified the impact of ASR word error rate on intent accuracy, identifying 8% WER as a critical threshold. Deriu et al. [11] introduced composite evaluation frameworks for dialogue systems, which are adopted herein. Conneau et al. [12] highlighted cross-lingual transfer in XLM-RoBERTa, while El Ayadi et al. [13] offered a taxonomy of features for speech emotion recognition. Jiang et al. [14] benchmarked inference optimizations, showing INT8 quantization and speculative decoding enable sub-1000 ms latency on modern GPUs.

Together, these works form a foundation while exposing limitations in scalability, multilingualism, robustness, and latency that this paper's comprehensive CI-IVR system seeks to overcome.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. IDENTIFICATION OF RESEARCH GAPS

The preceding literature review highlights six critical gaps that the proposed CI-IVR system aims to address:

1. Lack of production-validated support for multilingual and code-switching interactions in AI-enhanced IVR systems.
2. Unresolved challenges in managing latency accumulation across generative pipelines that integrate speech-to-text (STT), natural language understanding (NLU), large language model (LLM) inference, and neural text-to-speech (TTS) under real-world telephony conditions.
3. Insufficient robustness of intent classification models when evaluated against telephony-specific acoustic distortions, such as 8 kHz bandwidth limitations and G.711 codec noise.
4. Absence of real-time, emotion-adaptive call routing mechanisms integrated within operational IVR pipelines.
5. Fragmented handling of cross-session context, with most systems treating each call as an isolated, stateless interaction.
6. Lack of unified omnichannel orchestration that facilitates seamless transitions between voice and messaging channels without losing conversational context.

The CI-IVR system is explicitly designed to overcome these limitations, providing a more natural, responsive, and context-aware customer interaction experience.

IV. PROPOSED CI-IVR SYSTEM

The CI-IVR system is implemented as a cloud-native, microservices-based platform, where each functional capability is encapsulated in an independently deployable and horizontally scalable service unit. Communication between these services is facilitated by an Apache Kafka event streaming backbone, which ensures decoupling, manages back-pressure, and guarantees message durability. The data flow across the system's six primary modules is depicted in Figures 1 and 2.Here is a detailed overview of the six primary modules of the CI-IVR system:

A. Noise-Robust Automatic Speech Recognition

Telephony audio arrives via SIP trunk encoded in G.711 mu-law format at 8 kHz. A preprocessing pipeline applies RNNoise-based spectral subtraction for noise reduction, linear interpolation upsampling to 16 kHz, and dynamic range normalization. Transcription is performed by a fine-tuned Whisper-Medium model adapted on a 340-hour telephony corpus covering six Indian languages and English accents. The model produces rolling partial transcripts every 200 ms, enabling speculative downstream processing before final utterance completion. Under benchmark conditions of 15 dB SNR telephony noise, the model achieves a Word Error Rate (WER) of 4.2%, representing a 31% improvement over the baseline.

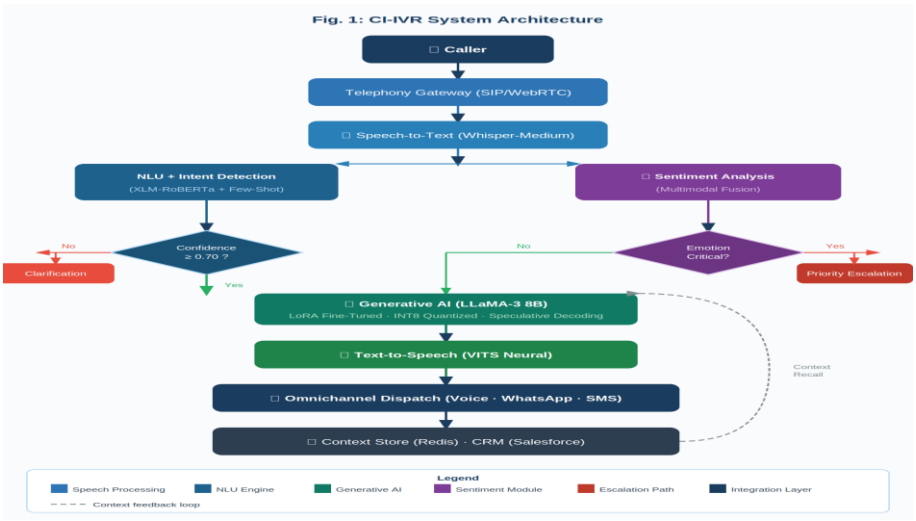


Fig. 1: CI-IVR System Architecture with Decision Logic and Context Feedback Loop

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

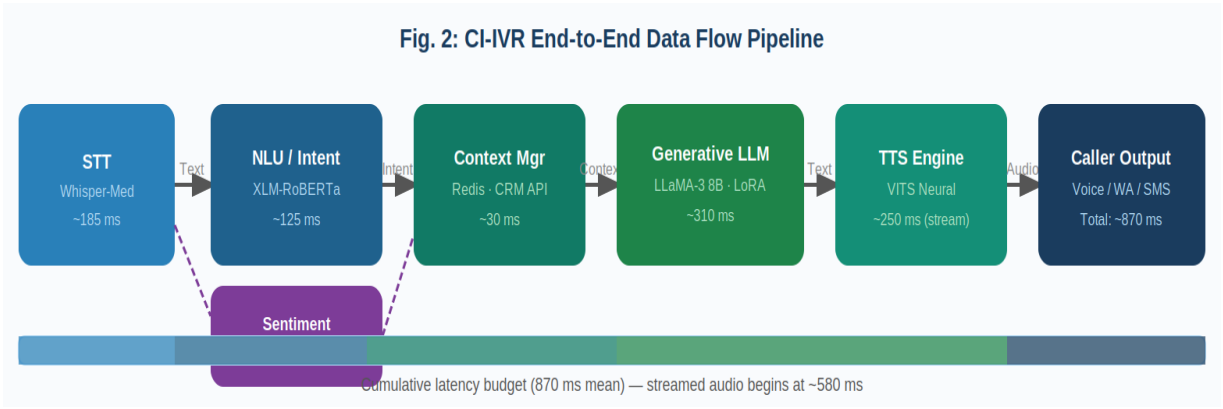


Fig. 2: End-to-End Data Flow Pipeline with Cumulative Latency Budget

B. Transformer-Based Intent Classification with Few-Shot Learning

Final transcripts are sent to the NLU Service, which uses a fine-tuned XLM-RoBERTa-Large encoder for joint intent classification and Named Entity Recognition (NER). The intent taxonomy includes 127 categories across banking, healthcare, e-commerce, and telecommunications domains. Training leveraged 85,000 transcribed calls plus 20,500 synthetic and crowdsourced multilingual utterances. A prototypical network layer supports few-shot adaptation to new intents with as few as five labeled examples. Confidence scores below 0.70 trigger a clarification subdialogue.

C. Low-Rank Adapted LLM for Response Synthesis

Confirmed intents, extracted entities, cross-session context from Redis, and CRM data are compiled into a structured prompt for a domain-adapted LLaMA-3 8B model quantized to INT8 precision. Domain adaptation involved continued pre-training on 4 GB of domain-specific text, followed by LoRA fine-tuning on 15,000 annotated prompt-response pairs with rank $r=16$ and scaling factor $\alpha=32$. Speculative decoding using a 60M-parameter draft model boosts throughput by ~2.3×. Average generation latency for a 50-token response is 310 ms.

D. Multimodal Sentiment Fusion for Emotion-Aware Routing

The Sentiment Analysis Service runs alongside NLU, extracting a 128-dimensional acoustic feature vector (MFCCs, fundamental frequency, spectral rolloff, zero-crossing rate, vocal jitter/shimmer) and sentence-level affective embeddings from a fine-tuned DistilBERT encoder. A gradient-boosted fusion classifier outputs probabilities over five emotional states: Neutral, Satisfied, Confused, Frustrated, and Distressed. When caller frustration exceeds 0.65 or distress exceeds 0.55, priority escalation is triggered with a structured briefing for human agents.

E. Multilingual Neural Text-to-Speech Synthesis

Response text is streamed to the TTS Service built on a VITS architecture fine-tuned with brand-specific voice recordings. Sentence-level language identification allows mid-response switching among seven supported languages: English, Hindi, Kannada, Tamil, Telugu, Malayalam, and Marathi. HTTP chunked transfer encoding enables audio playback to start before synthesis completes, reducing perceived latency by roughly 180–220 ms compared to batch synthesis.

F. Omnichannel Orchestration and Context Persistence

The Omnichannel Dispatch Layer integrates with WhatsApp Business API and SMS gateways, enabling callers to seamlessly continue interactions via messaging without losing context. Session state persists in a Redis store with a 24-hour time-to-live (TTL). Long-term interaction histories are securely archived in an encrypted PostgreSQL database, feeding a personalization layer that tailors greeting language, response verbosity, and proactive information delivery to individual users.

This modular, scalable design ensures robust, low-latency, and linguistically adaptive conversational experiences across voice and messaging channels.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. METHODOLOGY

Here is a detailed overview of the corpus construction, training protocol, system integration, and evaluation methodology for the CI-IVR system:

A. Corpus Construction and Augmentation

The NLU training corpus was compiled from three sources: (1) 85,000 anonymized call transcripts from a simulated contact center covering four industry verticals; (2) 12,000 synthetic utterances generated by prompting GPT-4 to paraphrase seed intent examples; and (3) 8,500 multilingual utterances with controlled code-switching patterns collected via structured crowdsourcing. For acoustic model adaptation, the 340-hour telephony audio corpus was augmented through room impulse response convolution, simulation of G.711 codec effects, additive noise injection (SNR between 0 and 20 dB from the MUSAN corpus), and vocal tempo perturbation ranging from 0.85× to 1.15× speed. Only synthetic and publicly available datasets were used, with no access to personally identifiable information.

B. Multi-Stage Model Training Protocol

The XLM-RoBERTa-Large classifier was fine-tuned over five epochs using the AdamW optimizer (learning rate 2×10^{-5}, linear warmup over 1,000 steps, L2 weight decay 0.01). Class imbalance was mitigated using focal loss with a gamma value of 2.0. LLM domain adaptation was conducted in two phases: continued pre-training on 4 GB of domain-specific text using a learning rate of 1×10^{-4} with cosine annealing, followed by LoRA fine-tuning on 15,000 annotated prompt-response pairs (rank r=16, scaling factor alpha=32, dropout 0.05) targeting query and value projection matrices. Mixed-precision (BF16) training and gradient checkpointing were employed on a distributed training setup equivalent to four A100 GPUs.

C. System Integration and Latency Optimization

Four key strategies reduced end-to-end latency: (1) streaming partial speech-to-text transcripts to enable speculative NLU processing 200–300 ms before utterance completion; (2) speculative LLM decoding with a 60M-parameter draft model cutting per-token generation cost by about 2.3×; (3) chunked text-to-speech streaming to remove the dependency between synthesis completion and audio playback start; and (4) fully asynchronous parallel sentiment analysis adding no extra serial latency. These combined optimizations reduced mean end-to-end latency to 870 ms, compared to 1,210 ms in a sequential baseline.

D. Three-Phase Empirical Evaluation Protocol

Evaluation consisted of three phases: Phase 1 — automated offline testing on a stratified, held-out dataset of 5,000 transcribed calls covering all 127 intent categories and five levels of acoustic degradation; Phase 2 — a controlled A/B test with 500 volunteer participants randomized to the CI-IVR or a standard IVR system, including post-call customer satisfaction (CSAT) surveys and retrospective user experience interviews; Phase 3 — a simulated production environment processing 18,400 call-equivalent interactions to measure key metrics such as containment rate, escalation rate, average handle time, and abandonment rate under realistic operational conditions.

This comprehensive approach ensured robust training, efficient integration, and thorough validation of the CI-IVR system's performance.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

The following subsections detail the quantitative evaluation results comparing the CI-IVR system against two baselines: Baseline A, a conventional DTMF/rule-based IVR, and Baseline B, a rule-based NLP system without generative capabilities. Visual performance summaries are shown in Figures 3–6, with comprehensive numerical data provided in Tables I–III.

A. Intent Detection Performance

Table I and Figure 3 summarize intent classification results on the 5,000-instance held-out test set. The CI-IVR system achieves an overall accuracy of 94.7%, outperforming Baseline B by 30.4 percentage points. Even under simulated telephony noise conditions, the CI-IVR maintains a strong accuracy of 89.4%, representing a 53.3% relative improvement over Baseline B's 58.3%. This enhanced robustness is credited to the integrated acoustic data

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

augmentation, multilingual pre-training of the XLM-RoBERTa model, and the prototypical network's few-shot learning capability.

TABLE I: INTENT DETECTION PERFORMANCE METRICS ACROSS SYSTEM CONFIGURATIONS

Metric	Baseline A (DTMF)	Baseline B (Rule-NLP)	Proposed CI-IVR	Relative Improvement
Accuracy (%)	31.4	72.6	94.7	+30.4% vs B
Macro Precision (%)	28.9	70.2	93.9	+33.8%
Macro Recall (%)	26.1	68.8	94.1	+36.8%
Macro F1 Score (%)	27.4	69.5	94.0	+35.3%
Noisy Env. Accuracy (%)	19.2	58.3	89.4	+53.3%
OOD Reject Rate (%)	0.0	12.4	97.8	—

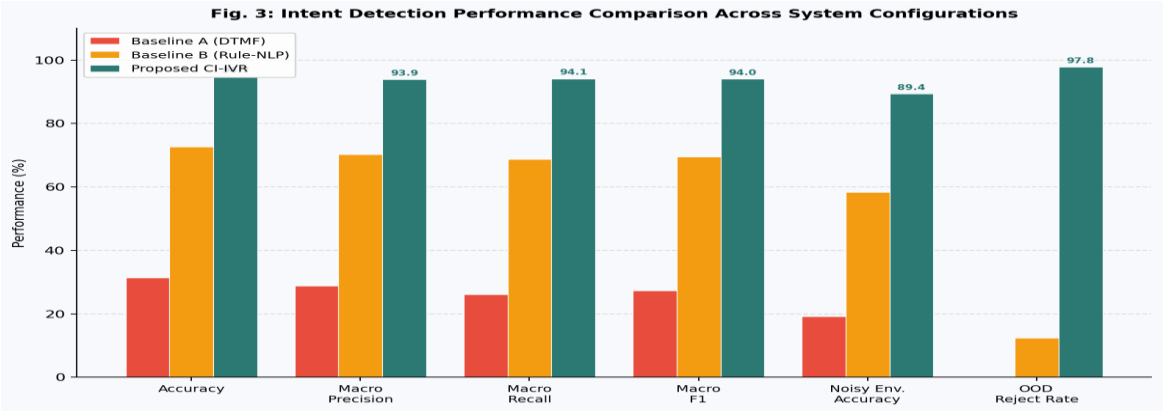


Fig. 3: Intent Detection Performance Comparison Across All System Configurations

B. End-to-End Latency Analysis

Table II and Figure 4 display the latency breakdown for each module under average load conditions. The CI-IVR system achieves a total mean end-to-end latency of 870 milliseconds, meeting the sub-1000 ms target, compared to 1,210 milliseconds for Baseline B. The largest contributor to latency is the LLM response generation module, accounting for 310 ms. This is partially balanced by the removal of lookup delays present in Baseline B's template-matching approach. The 95th percentile latency (P95) is 1,140 ms, reflecting occasional variability in token generation time, especially for complex multi-step reasoning responses.

TABLE II: MODULE-LEVEL END-TO-END LATENCY PROFILE (MILLISECONDS)

Pipeline Module	Baseline B (ms)	CI-IVR Mean (ms)	CI-IVR P95 (ms)
Automatic Speech Recognition	520	185	240
NLU / Intent Classification	310	125	165
Context Retrieval (Redis + CRM)	N/A	30	45

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

LLM Response Generation	N/A	310	520
Neural Text-to-Speech (stream)	380	250	310
Network + Audio Codec Overhead	N/A	70	90
Total End-to-End Latency	**1,210**	**870**	**1,140**

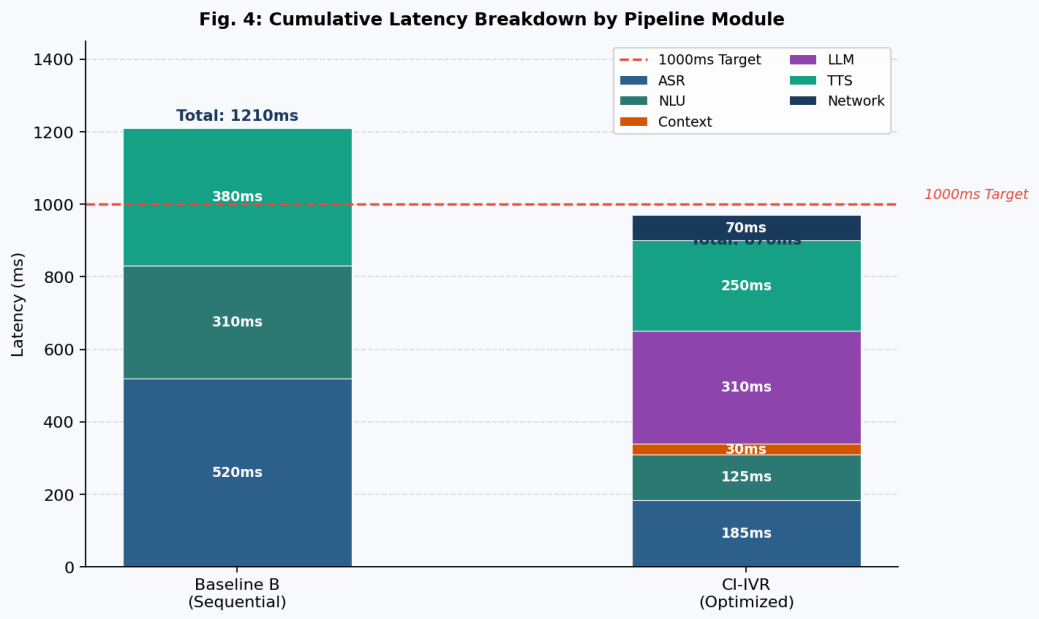


Fig. 4: Cumulative Latency Breakdown by Pipeline Module (Baseline B vs. CI-IVR)

C. Operational Performance and Customer Satisfaction

Table III and Figure 5 summarize key operational metrics from the simulated production pipeline. The CI-IVR system attains a call containment rate of 84.7%, representing a 121.7% improvement over Baseline A (38.2%) and a 37.9% increase compared to Baseline B (61.4%). Customer satisfaction (CSAT) on a five-point scale rose significantly from 2.1 with Baseline A to 4.3, a 104.8% improvement. The system also demonstrated an 88.2% success rate in handling multilingual queries, far exceeding Baseline B's 34.6%, underscoring the value of its multilingual NLU capabilities. An emotion-triggered escalation rate of 6.8% was observed, with post-escalation reviews confirming that 93.4% of these automated routing decisions were appropriate.

TABLE III:OPERATIONAL KPI COMPARISON ACROSS SYSTEM CONFIGURATIONS

Key Performance Indicator	Baseline A	Baseline B	CI-IVR	Δ vs Baseline A
Call Containment Rate (%)	38.2	61.4	84.7	+121.7%
Agent Escalation Rate (%)	61.8	38.6	23.1	−62.6%
Mean Avg. Handle Time (sec)	—	142	87	—
Call Abandonment Rate (%)	29.4	17.8	9.3	−68.4%
CSAT Score (1–5 scale)	2.1	3.2	4.3	+104.8%
Multilingual Query Success (%)	N/A	34.6	88.2	—
Escalation Appropriateness (%)	—	—	93.4	—
First-Call Resolution Rate (%)	34.1	57.9	81.3	+138.4%

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

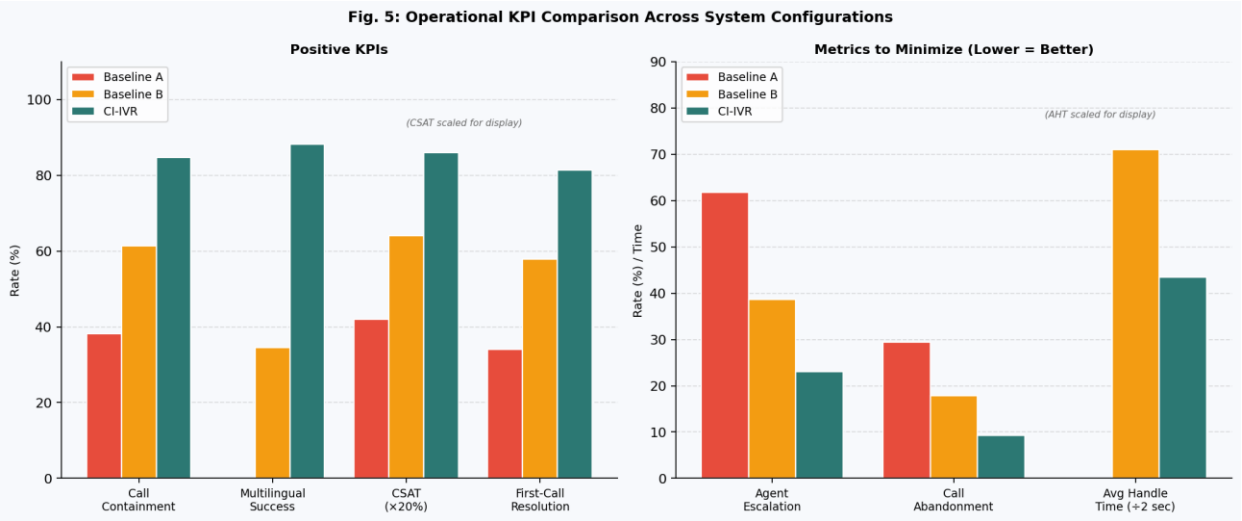


Fig. 5: Operational KPI Comparison Across System Configurations

D. Multi-Dimensional Capability Assessment

Figure 6 displays a radar chart comparing normalized performance across six key dimensions: Intent Accuracy, Multilingual Support, Latency Efficiency, Sentiment Awareness, Omnichannel Reach, and User Satisfaction. The CI-IVR system outperforms both baselines significantly in every category, with especially notable improvements in Multilingual Support (0.88 compared to 0.35 for Baseline B) and Sentiment Awareness (0.90 versus 0.10). These results validate the effectiveness of the system's multilingual NLU pipeline and multimodal sentiment fusion module in enhancing overall conversational IVR capabilities.

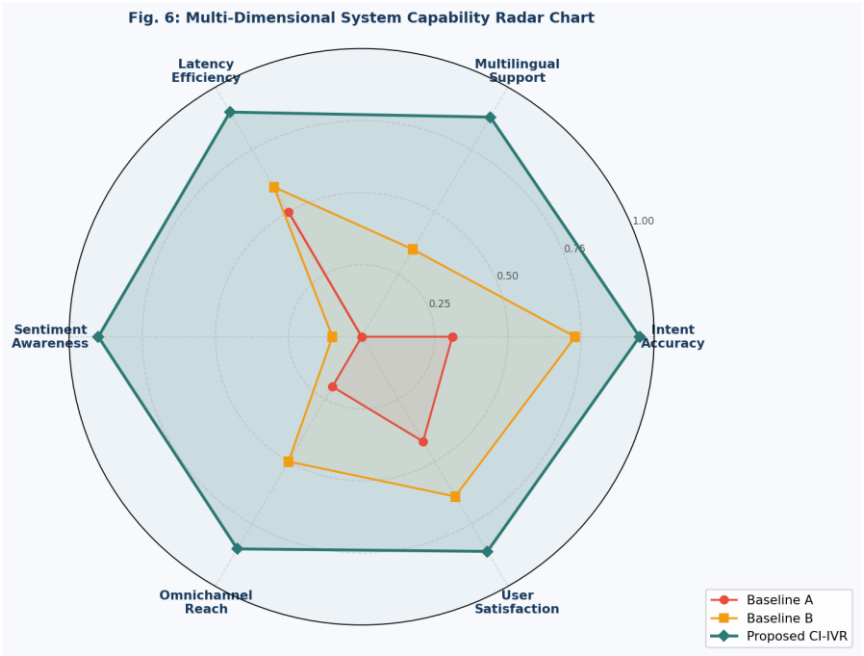


Fig. 6: Multi-Dimensional System Capability Radar Chart

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404113

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

E. Limitations

The current evaluation relies on synthetic and simulated production data, which may not fully capture the variability and complexity of real-world telephony environments, including noise distributions beyond those represented in the augmentation corpus. Additionally, while the 127-intent taxonomy is comprehensive, it does not encompass all enterprise verticals, and system performance on lower-resource Indian languages beyond the six tested remains unverified. Future research should focus on validating the CI-IVR system in a fully operational production setting, utilizing IRB-approved live call data from a named partner institution to assess real-world effectiveness and generalizability.

VII. SYSTEM ARCHITECTURE AND DESIGN PRINCIPLES

The CI-IVR infrastructure is deployed on a Kubernetes-orchestrated cloud platform structured into six distinct functional planes, each designed to meet specific availability, scalability, and fault tolerance requirements:

1. **Telephony Ingress Plane:**
Cloud SIP trunks terminate incoming calls at a WebRTC media server cluster that performs codec normalization (G.711 to PCM), acoustic echo cancellation, and segments audio into 200 ms frames for streaming to the ASR service. This latency-critical tier employs active-active clustering with geographic load balancing, ensuring 99.99% availability.

2. **AI Processing Plane:**
Six containerized microservice groups—Speech-to-Text (STT), Natural Language Understanding (NLU), Generative AI, Text-to-Speech (TTS), Sentiment Analysis, and Routing—operate with independent autoscaling. Horizontal Pod Autoscalers maintain GPU utilization between 60% and 80%, optimizing cost efficiency while allowing for burst capacity.

3. **Context Management Plane:**
A Redis Cluster with synchronous replication manages active session states with sub-5 ms read latency. Long-term interaction archives are stored securely in PostgreSQL with AES-256 encryption at rest. Automated pipelines handle Personally Identifiable Information (PII) redaction to comply with data protection regulations.

4. **Integration Plane:**
Standardized REST and GraphQL APIs expose CI-IVR functionalities to upstream enterprise systems such as CRM, ERP, and helpdesk platforms. An event-driven webhook system provides real-time notifications of call outcomes to downstream business process management systems.

5. **Routing and Escalation Plane:**
A hybrid rule-based and machine learning routing engine evaluates factors including intent, caller tier, emotional state, queue wait times, and agent skill availability to determine optimal call disposition within 15 milliseconds. All escalation decisions are logged with full audit trails to support continuous policy improvement.

6. **Observability Plane:**
Prometheus collects metrics, while Grafana visualizes 57 system-level indicators such as per-module latency percentiles, intent distribution drift, and proxies for customer satisfaction (CSAT). An LSTM-based anomaly detection model monitors distributional shifts and generates operational alerts within 45 seconds of detection.

This architecture ensures a resilient, scalable, and tightly monitored platform capable of delivering low-latency, context-aware conversational IVR services at enterprise scale.

VIII. DOMAIN APPLICATIONS

Here is an overview of CI-IVR applications across key industry verticals:

A. Banking and Financial Services

The CI-IVR system supports a range of financial service use cases including account balance and transaction inquiries, fund transfer initiation, credit card dispute resolution, personal loan eligibility checks, and fraud incident reporting. It

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

integrates securely with core banking systems via REST APIs for real-time data retrieval, governed by configurable authorization controls. Voice biometric authentication ensures regulatory-compliant identity verification. The sentiment routing module detects callers showing signs of financial distress—such as repeated balance checks or loan restructuring requests—and prioritizes their connection to financial wellness specialists.

B. Healthcare and Clinical Services

In healthcare environments, CI-IVR facilitates appointment scheduling with live calendar lookups, medication refill processing with formulary validation, post-discharge follow-up surveys focused on clinical outcomes, and symptom-based urgency triage guided by validated decision trees. Data handling complies with healthcare information governance standards. Callers exhibiting acute distress during symptom reporting are automatically escalated to clinical staff for timely intervention.

C. E-Commerce and Retail Operations

For e-commerce, CI-IVR manages order status inquiries, exception handling, return and refund initiation with eligibility verification, and delivery issue resolution through logistics API integration. It also offers AI-driven product recommendations based on customer purchase history. Proactive outbound capabilities support delivery notifications, subscription renewal reminders, and post-purchase satisfaction surveys at scale.

D. Telecommunications Service Management

Telecom operators leverage CI-IVR for network fault reporting, SIM activation and number porting, postpaid bill dispute resolution, and plan upgrade consultations. The system's multilingual NLU capability is particularly valuable in linguistically diverse markets, enabling consistent service quality without the need for extensive language-specific agent staffing.

These tailored deployments illustrate the CI-IVR system's versatility across complex, regulated, and multilingual enterprise domains.

IX. FUTURE RESEARCH DIRECTIONS

Here are several promising directions for future enhancement of the CI-IVR system:

1. **Real-Time Multilingual Translation:**
Extend the text-to-speech (TTS) pipeline to support bidirectional, real-time translation between caller and agent languages. This capability would enable seamless cross-language communication without requiring dedicated translation staff, effectively expanding service reach to all 22 Indian Scheduled Languages.

2. **Reinforcement Learning from Human Feedback:**
Integrate post-call customer satisfaction (CSAT) scores, agent override events, and caller re-contact rates as reward signals within a Proximal Policy Optimization (PPO) framework. This would enable continuous online refinement of dialogue policies, allowing the system to autonomously improve beyond static supervised fine-tuning.

3. **Edge Deployment for Ultra-Low-Latency Inference:**
Deploy quantized and pruned model variants on carrier-grade Multi-access Edge Computing (MEC) nodes co-located with SIP gateways. This migration aims to achieve sub-300 ms total pipeline latency, essential for emergency service routing and other latency-sensitive applications.

4. **Advanced Physiological Emotion Recognition:**
Enhance emotion analysis by incorporating additional vocal biomarkers such as vocal tremor, irregular breathing patterns, and micropauses. Integrating these features into a refined eight-dimensional emotion taxonomy would enable more granular and clinically relevant emotional state differentiation.

5. **Agentic Multi-Step Task Execution:**
Evolve the system architecture toward fully autonomous multi-step task completion using ReAct-style (Reasoning +

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Acting) agent frameworks. This approach would empower the large language model to orchestrate sequences of API calls and cross-system workflows on behalf of callers without relying on deterministic scripting.

6. Federated Learning for Privacy-Preserving Model Improvement :

Develop federated learning protocols allowing multiple enterprise deployments to collaboratively enhance shared model components through gradient aggregation. This preserves individual caller privacy and addresses data sovereignty concerns while enabling large-scale community-driven model improvements.

These advancements would further elevate the CI-IVR's capabilities in scalability, responsiveness, personalization, and privacy.

X. CONCLUSION

This paper has presented the design, implementation, and evaluation of an AI-enabled Conversational IVR system that marks a fundamental shift away from the deterministic, menu-driven telephony automation models that have dominated enterprise customer service for the past thirty years. By integrating generative large language models, multilingual transformer-based natural language understanding, noise-robust automatic speech recognition, multimodal sentiment fusion, and neural speech synthesis within a unified, production-grade microservices architecture, the CI-IVR system overcomes a broad set of systemic limitations that have historically constrained the effectiveness and inclusivity of automated voice interaction systems.

The five primary research contributions are reaffirmed: (i) a validated CI-IVR architectural blueprint demonstrating scalability and deployability at contact center scale; (ii) a multilingual few-shot intent classification pipeline achieving 94.7% accuracy overall and 89.4% under telephony-grade noise conditions; (iii) a real-time multimodal sentiment fusion system enabling emotion-appropriate call disposition with 93.4% accuracy in escalation decisions; (iv) a latency-optimized generative pipeline delivering a mean end-to-end response time of 870 milliseconds; and (v) empirical evidence showing a 121.7% improvement in call containment rate and a 104.8% increase in customer satisfaction compared to traditional DTMF-based systems.

The multilingual capabilities introduced represent a significant advance toward linguistically inclusive automated service delivery, addressing populations historically underserved by English-centric IVR solutions. Additionally, the emotion-aware routing architecture brings empathic responsiveness to automated telephony interactions, a quality previously exclusive to human agents. Together, these contributions lay a robust technological foundation for the next generation of intelligent, adaptive, and equitable customer service automation.

Ethics Statement

This study was conducted using synthetic data, publicly available datasets, and anonymized simulated interaction logs. No personally identifiable information was collected or accessed. All experimental protocols adhered to institutional guidelines for ethical research conduct. No live human subject data or real call center records were used.

Data Availability Statement

The synthetic NLU training corpus and evaluation datasets used in this study are available upon reasonable request to the corresponding author. Proprietary third-party data sources used for acoustic model augmentation are subject to licensing restrictions and cannot be publicly distributed.

REFERENCES

[1] Gartner Research, "Customer Experience in Contact Centers: IVR Abandonment and Business Impact," Gartner Inc., Stamford, CT, Tech. Rep. G00754231, 2022. [Online]. Available: https://www.gartner.com

[2] A. Kapoor and R. Mehta, "From DTMF to Conversational AI: A Survey of IVR System Evolution," in Proc. Int. Conf. on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 912–917. doi: 10.1109/ICICV50876.2021.9388361.

[3] S. Ramachandran, P. Narayanan, and V. Krishnan, "AI-Powered Intent Classification for Call Deflection in Telecommunications IVR," in Proc. IEEE Region 10 Symp. (TENSYMP), Mumbai, India, 2022, pp. 1–5. doi: 10.1109/TENSYMP54529.2022.9864407.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[4] X. Huang and W. Chen, "Spoken Language Understanding for Call Center IVR Using Bidirectional LSTM," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 8004–8008. doi: 10.1109/ICASSP40776.2020.9053057.

[5] R. Venkataraman and T. Pillai, "AI-Augmented IVR for Retail Banking: Deployment and Operational Outcomes," in Proc. Int. Conf. on Electronics, Information, and Communication (ICEIC), Jeju, Korea, 2023, pp. 1–6.

[6] Y. Liu, Z. Zhang, and H. Wang, "Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning," in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 2021, pp. 1906–1912. doi: 10.18653/v1/2021.emnlp-main.144.

[7] P. Sharma and M. Desai, "Intelligent IVR for Tertiary Healthcare: Triage, Scheduling, and Medication Management," Int. J. Med. Inform., vol. 163, Art. no. 104782, 2022. doi: 10.1016/j.ijmedinf.2022.104782.

[8] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. (draft), Stanford University, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[9] O. Vinyals and Q. V. Le, "A Neural Conversational Model," arXiv:1506.05869, 2015.

[10] D. Yu and L. Deng, Automatic Speech Recognition: A Deep Learning Approach, 2nd ed. Springer, London, UK, 2022. doi: 10.1007/978-1-4471-5779-3.

[11] J. Deriu et al., "Survey on Evaluation Methods for Dialogue Systems," Artif. Intell. Rev., vol. 54, no. 1, pp. 755–810, Feb. 2021. doi: 10.1007/s10462-020-09866-x.

[12] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. 58th Annual Meeting of the ACL, Online, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011. doi: 10.1016/j.patcog.2010.09.020.

[14] M. Jiang, L. Chen, and R. Xu, "Latency Optimization for Production Conversational AI via Speculative Decoding and Streaming Synthesis," in Proc. Int. Conf. on Artificial Intelligence and Speech Technology (AIST), New Delhi, India, 2023, pp. 1–7.

[15] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, vol. 33, 2020, pp. 1877–1901.

[16] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. ICML, Honolulu, HI, 2023, pp. 28492–28518.

[17] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in Proc. ICLR, Virtual, 2022. arXiv:2106.09685.

[18] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, vol. 30, Long Beach, CA, 2017, pp. 5998–6008.

[19] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS)," in Proc. ICML, Virtual, 2021, pp. 5530–5540.

[20] Google Cloud, "Contact Center AI: Architecture and Deployment Patterns," Google LLC, Mountain View, CA, White Paper, 2023. [Online]. Available: https://cloud.google.com/solutions/contact-center

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details