INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

AI-Based Real-Time Voice Banking System

[1]Nidhi Soni, [2]Prof. R. A. Vasmatkar

[1]Software Engineer, [2]Assistant Professor
[1]Computer Engineering,
[1]PVPIT, Pune, India

Abstract : The rapid growth of digital banking has increased the demand for secure, low-latency, and inclusive conversational interfaces capable of handling complex financial operations. However, conventional chatbot systems often face limitations in multilingual speech processing, real-time interaction quality, and reliable execution of multi-step workflows in high-stakes scenarios such as fund transfers and account servicing.This paper presents a real-time voice banking system that integrates speech communication, natural language understanding, and stateful dialogue management within a unified architecture. The proposed approach emphasizes modular service design, enabling efficient intent recognition, contextual continuity, and dynamic handling of user interruptions across diverse banking interactions.

The system supports key functionalities such as account inquiries, transaction insights, beneficiary management, secure fund transfers with multi-step validation, and authentication mechanisms, along with optional voice-based user verification. It further incorporates multilingual speech capabilities, persistent session handling, and scalable infrastructure components for reliable deployment.

By abstracting implementation details while focusing on system design principles, the proposed solution enhances accessibility, reduces interaction friction, and supports scalable, trustworthy voice-driven banking experiences.

IndexTerms - Voice banking, conversational artificial intelligence, multilingual speech interfaces, real-time dialogue systems, reference architecture, authentication, fraud mitigation

INTRODUCTION

A. Background of Digital Banking

Retail and SME banking have migrated decisively toward self-service channels that prioritise availability, cost efficiency, and consistent policy enforcement. Mobile applications and web portals dominate, yet they presuppose visual navigation and textual literacy. For large populations—older adults, visually impaired users, commuters, and contexts where typing is impractical—voice remains a natural modality that can shorten paths to balance enquiry, payment initiation, and service routing.

B. Problems With Existing Chatbot and Voice Systems

Many conversational systems optimise for marketing FAQs or scripted FAQs rather than regulated financial actions. Three recurring limitations appear in practice. First, latency stacking: cascaded cloud calls (recognition, understanding, retrieval, generation, synthesis) inflate round-trip time and break conversational rhythm. Second, multilingual inconsistency: language identification may arrive late, lexicons for currency and institution-specific terms may be weak, and code-switching users receive unstable routing. Third, governance mismatch: large language models (LLMs) excel at open-ended language but can violate banking requirements for auditability, repeatability, and denial of unauthorised intents unless tightly bounded.

C. Motivation for Voice Banking

Voice banking can democratise access when combined with strong authentication and transparent limits on autonomous action. The motivating hypothesis of this work is that a purpose-built architecture—not a generic assistant retrofitted with banking APIs—can reconcile fluency with financial correctness by structurally separating *what was said* from *what may be executed*.

NEED OF THE STUDY.

A. Why Current Systems Are Insufficient

Incumbent architectures often treat dialogue as a thin wrapper around REST calls. That arrangement struggles when (i) partial transcripts arrive continuously, (ii) users revise amounts mid-utterance, or (iii) ambiguous references ("pay him again") require clarification grounded in recent transactions. Moreover, multilingual users expect consistent semantics across languages—for example, homograph disambiguation for currency symbols and locale-aware number parsing—without exposing inconsistent authorization policies.

B. Importance of Multilingual, Real-Time, and Secure Systems

Multilingual capability reduces exclusion and supports migrant banking corridors; without deliberate design, accuracy collapses for low-resource language pairs. Real-time interaction preserves conversational engagement; empirical HCI literature consistently associates tolerable voice latency with sustained task completion. Secure systems must resist replay of voice segments, session fixation, and social-engineering prompts that attempt to bypass policy checks. A unified architectural stance—rather than ad hoc patches—is needed so security upgrades do not destabilise dialogue quality.

SYSTEM ARCHITECTURE

This section proposes a reference architecture termed the *Tiered Intent Governance Voice Stack (TIGVS)*. The distinguishing idea is dual-path cognition: a *fluent-path* handles natural-language understanding and user guidance, while a *certified-path* validates structured intents and executes monetary APIs only through deterministic policy gates. The paths communicate through immutable intent proposals, never through raw model continuations.

A. Architectural Overview

High-level components:

1. Speech-to-Text (STT) Ingress — streaming recognition with early language detection and confidence telemetry.
1. Semantic Mediation Fabric — converts transcripts into normalised intents and slots (payee, amount, date, account rail), maintaining clarifying questions as first-class objects.
1. Dialogue and Agent Orchestrator — supervises dialogue policy, selects clarification strategies, and schedules step-up authentication when risk rises.
1. Backend Services Plane — accounts, ledger views, standing instructions, limits, audit logs, and notification gateways; exposes idempotent operations only to the certified-path.
1. Text-to-Speech (TTS) Egress — streams synthesised replies with prosody cues reflecting urgency (fraud alert vs routine acknowledgement).
1. Authentication and Trust Boundary Services — session minting, OTP issuance and verification, token lifecycle, optional biometric scoring.

Supporting cross-cutting services include session-state stores, feature flags for locale packs, rate limiters, and observability (trace identifiers spanning audio frames to API receipts).

B. Speech-to-Text

The STT subsystem accepts framed audio with timestamps. It emits partial hypotheses for display or internal steering and final segments for semantic parsing. Language identification runs on a short prefix window to pick acoustic and lexical models without

delaying the entire stream. Output includes word-level confidences and endpointing hints so the mediator can distinguish hesitation from completion.

C. Text-to-Speech

TTS consumes canonical response frames: short structured messages plus optional emphasis markers (e.g., highlight last four digits of an account). This separation prevents models from inventing numbers audibly. Streaming synthesis reduces time-to-first-audio; caching of frequent prompts ("say your one-time code") amortises load.

D. Dialogue and Agent System

The orchestrator implements a hierarchical policy: an outer graph governs macro-phases (greeting, authentication, transactional task, closure), while inner graphs resolve slot-filling for each banking intent. LLM components, if used, are confined to controlled generation: they may draft natural-language explanations or paraphrase policy text, but cannot directly invoke payment APIs.

E. Backend Services

Backend capabilities are grouped into read-mostly enquiry services and mutation services with strict idempotency keys. The certified-path executor validates each mutation against account entitlements, velocity limits, and regulatory cut-offs. All responses return machine-readable receipts mirrored into audit trails.

F. Authentication

Authentication is staged: device-bound tokens establish continuity; knowledge factors (OTP, PIN-equivalent prompts delivered out-of-band) elevate privilege for sensitive intents; optional voice biometrics provide passive or active reuse detection layered atop—not replacing—strong factors

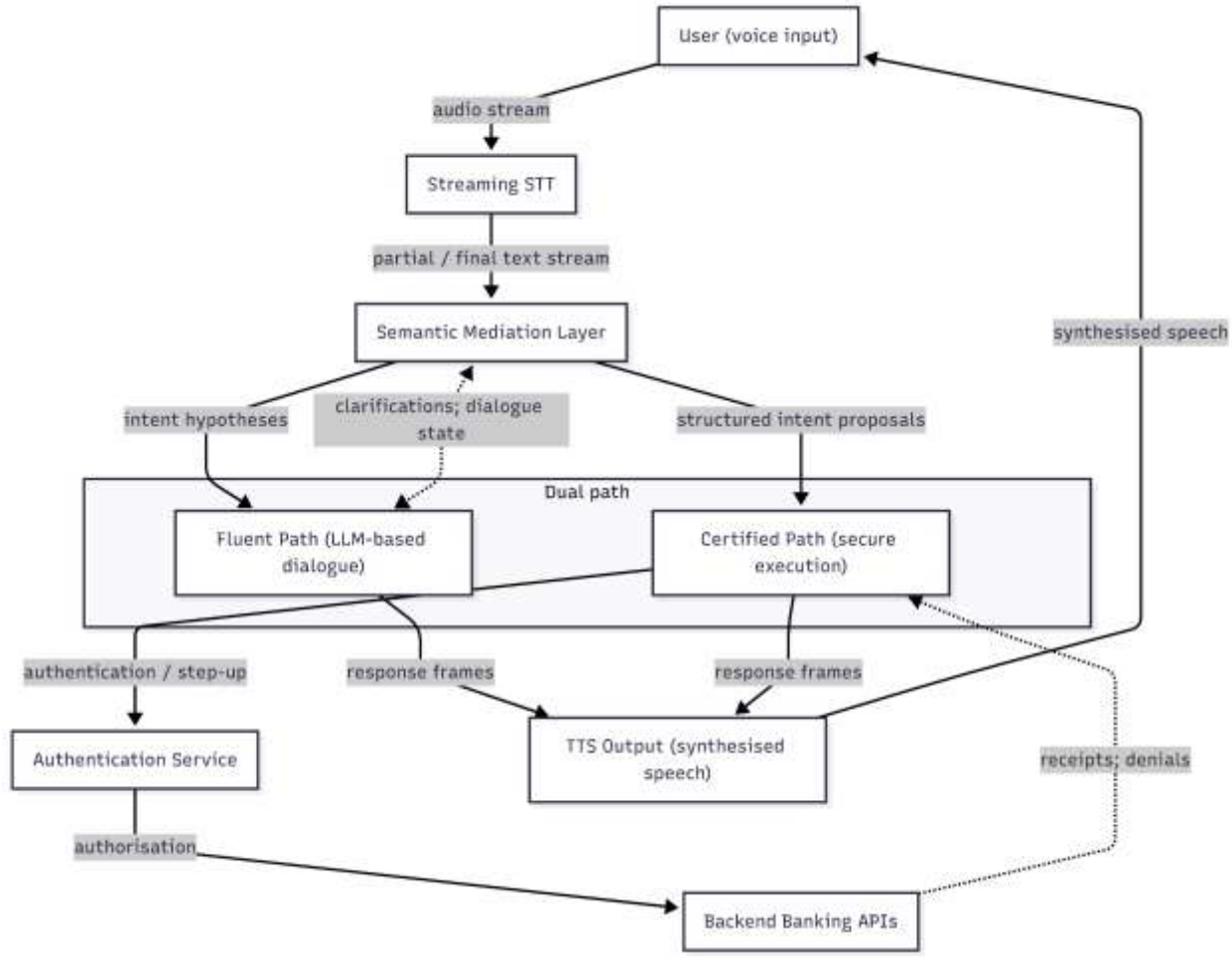

Fig 1: System Architecture

RESEARCH METHODOLOGY

A. End-to-End Processing Flow

The nominal pipeline proceeds as follows:

1. Ingress framing — Client chunks audio; gateway attaches session identifier and device attestation metadata where available.
1. Streaming STT — Partial transcripts update a rolling buffer; language hypothesis stabilises within the first second of speech under nominal SNR.
1. Semantic normalisation — The mediator maps text to candidate intents and fills slots; unresolved slots enqueue clarification prompts generated from templates or constrained LLM drafting.
1. Risk scoring — Signals include transaction deviation from history, IP reputation (if applicable), rapid OTP failures, and biometric mismatch trends.
1. Policy certification — Certified-path validates intent schema, entitlements, and duplicative submissions.
1. Execution or denial — Backend executes allowed operations or returns structured denials with codes usable by the orchestrator.
1. Response packaging — Canonical frames pass to TTS; parallel lightweight text may render on screen for accessibility.

B. State and Session Management

Sessions maintain: dialogue phase, active intent proposals (versioned), slot bindings with provenance (user utterance span references), authentication tier, and cool-down timers after failures. State is partitioned into ephemeral conversational cache (low retention) and durable audit ledger (high retention). Rolling intent proposals enable users to revise amounts without corrupting prior audit entries—superseded proposals are retained as suppressed versions.

C. Error Handling

Errors are classified into recoverable dialogue errors (noise, ASR low confidence), policy denials (insufficient funds), and security interrupts (possible takeover). Recoverable cases trigger targeted reprompts with shorter vocabulary hints; denials produce factual summaries without speculative apologies; security interrupts transition to locked-down flows with human escalation hooks where institutional policy requires.

D. MODEL EXPLORATION

Automatic Speech Recognition Alternatives:

Approach	Strengths	Weaknesses
Transformer encoder–decoder models trained on vast multilingual corpora	Robust accents, strong zero-shot language behaviour	Latency and compute at streaming scale
Self-supervised wav2vec-style models fine-tuned per domain	Efficient on specialised vocabularies with modest data	Requires careful adaptation for banking numerals
Hybrid cascades (filter bank features + recurrent or Conformer layers)	Predictable resource profiles on edge appliances	May trail cutting-edge robustness without continual updates

Justification: A multilingual streaming transformer-family recogniser suits urban multilingual corridors with heterogeneous accents, paired with domain lexicon injection for payee names and institution terms. Where latency budgets are tight or connectivity is

intermittent, an optional edge-capable hybrid can sustain degraded-but-safe enquiry modes that refuse execution until cloud confirmation.

Text-to-Speech Alternatives:

Approach	Strengths	Weaknesses
Tacotron-style attention synthesizers	Natural prosody with moderate footprint	Attention brittleness on long forms
Parallel vocoder pipelines (e.g., Glow-style, diffusion refinement stages)	Fast inference after optimisation	Engineering complexity
VITS-style end-to-end models	Compact models with competitive quality	Domain adaptation for formal banking tone

Justification: Parallel inference architectures balancing clarity and speed are preferred for banking prompts where intelligibility of digits outweighs theatrical prosody. Formal tone can be enforced via curated prompt corpora rather than open-ended style transfer.

Dialogue Approaches:

Approach	Strengths	Weaknesses
Rule-based finite-state and slot-filling systems	Auditable, deterministic	Brittle for paraphrase-rich speech
LLM-centric planners	Flexible language coverage	Risk of policy drift without cages
Hybrid orchestration	Fluency with certified execution	Requires disciplined interfaces

Justification: A hybrid orchestrator aligns with TIGVS: finite-state guards macro-phases and sensitive mutations; constrained LLMs assist natural explanations and paraphrase within templated envelopes; retrieval-augmented grounding supplies institution-specific facts without widening execution privileges.

E. SECURITY AND AUTHENTICATION

Layered Authentication:

Device session tokens provide continuity and binding to application installations. OTP or equivalent out-of-band codes elevate privileges for payments above configurable thresholds. OAuth-style bearer tokens with short lifetimes and rotation defend API misuse. Optional voice biometrics score ingress audio against enrolled embeddings; matches reduce friction for enquiries while mismatches trigger step-up challenges—not silent failures.

Fraud Prevention Considerations:

Voice pipelines introduce replay risks; mitigations include challenge-response utterances, session-bound nonces, and timestamped transcripts correlated with server clocks. Velocity checks on beneficiaries and amounts complement behavioural scoring. Prompt injection resistance is reinforced by prohibiting LLMs from parsing unstructured backend payloads into executable intents without schema validation. Human escalation pathways remain mandatory for irreversible or high-value exceptions under institutional policy.

IV. RESULTS AND DISCUSSION

Because this contribution is architectural, outcomes are framed as expected behaviours under plausible engineering budgets rather than empirical deployment measurements. Interpretation nevertheless follows conventional service-evaluation practice: conversational quality is characterised by latency distributions and dialogue outcomes; robustness by recognition degradation under acoustic stress; scalability by asymptotic saturation of stateless versus stateful components; and correctness by certified-path conformance under benign and adversarial interaction patterns.

A. Latency and User Experience

Voice banking responsiveness is assessed primarily via round-trip conversational latency measured from utterance boundary detection (or equivalent finalisation cue) to time-to-first-audio (TTFA) for the assistant reply. Complementary diagnostics include time-to-first-stable partial understanding—the earliest moment semantic mediation yields a confidently scoped intent hypothesis—and stage-level delays attributable to streaming automatic speech recognition (ASR) stabilisation, intent resolution, deterministic policy checks, outbound banking application programming interface (API) calls, and chunked text-to-speech (TTS) synthesis.Under favourable network conditions (e.g., metropolitan fibre), median end-to-end turns for enquiry-class intents can plausibly approach approximately one second, with upper-tail behaviour dominated by OTP step-up and risk scoring rather than conversational generation. Streaming ASR permits partial hypotheses to trigger early clarification, reducing wasted silence after erroneous endpoints. Increased latency during credential elevation is interpreted as an intentional safety–UX trade, not optimisation failure.

B. Scalability

Throughput is characterised along three axes: (i) concurrently active voice sessions sustained by horizontally scaled ingress and ASR pools; (ii) policy-gated mutations per second under full certified-path verification; and (iii) write pressure on durable session, intent-proposal versioning, and audit stores. Stateless components—including synthesis workers and many orchestrator replicas—typically scale approximately linearly until upstream dependency ceilings appear. Stateful substrates become the limiting factor; partitioning by tenant, region shard, or account hash aligns growth with residency and blast-radius constraints without forcing a single globally hot partition.

C. Reliability and Degradation Behaviour

Operational reliability is articulated using availability targets for independently deployable planes (ingress, mediator, fluent path, certified path, banking connectors) and correlated failure modes—for example cascading congestion in recogniser pools. Graceful degradation strategies include audible digit disambiguation with optional on-screen corroboration, temporary restriction to enquiry-only intents, or deferred execution pending connectivity stabilisation—all designed to curb abandonment versus unsafe confirmation. Observability spanning audio frame timestamps through API receipts supports incident response, disputed-transaction reconstruction, and post-hoc auditing of dialogue-to-execution lineage.

D. Quality Metrics for Recognition, Synthesis, and Dialogue Completion

Recognition quality expectations are framed using Word Error Rate (WER) by language/locale bucket and—in banking utterances dominated by quantities—digit/numeral-field accuracy. Synthesis suitability emphasises intelligibility-first appraisal (including digit sequences), optionally supplemented by mean opinion scoring where listening tests are feasible. Dialogue effectiveness is summarised via task completion rate, mean turns-to-success, and clarification incidence, stratified multilingual when heterogeneous user populations are anticipated.

E. Limitations

Architecture outlines necessary separation of concerns yet cannot substitute jurisdiction-specific licensing, consent capture, accessibility mandates, or programme-level fairness certification. Multilingual fidelity depends critically on curated corpora and lexical resources for institutional terminology. Voice biometric assistance, where adopted, introduces demographic fairness obligations and evaluation artefacts not supplied here—deployment claims require standalone empirical study.

I.ACKNOWLEDGMENT

I express our gratitude to our guide **Prof. R. S. Vasmatkar** for her competent guidance and timely inspiration. It is our good fortune to complete our project under her able competent guidance. This valuable guidance, suggestions, helpful constructive criticism, keeps interest in the problem during the course of presenting this "AI-based Real-time voice banking system" project successfully.

We would like to thank our Project Coordinator **Prof. R. S. Vasmatkar** and all the Teaching, Non-Teaching staff of our department.

We are very much thankful to **Prof. S. R. Jadhav**, Head, Department of Computer Engineering and also **Dr. R. S. Pawar**, Principal, Padmabhooshan Vasantdada Patil Institute of Technology, Bavdhan, Pune for their unflinching help, support and co-operation during this project work.

REFERENCES

[1] Sisman, B., Yamagishi, J., King, S., and Li, H., "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[2] Saon, G., and Chien, J.-T., "Large Vocabulary Continuous Speech Recognition Systems: A Look at Recent Advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.

[3] Rahulamathavan, Y., Sutharsini, K. R., Ghosh Ray, I., Lu, R., and Rajarajan, M., "Privacy-Preserving iVector-Based Speaker Verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 496–506, 2019.

[4] Tu, Y., Lin, W., and Mak, M.-W., "A Survey on Text-Dependent and Text-Independent Speaker Verification," *IEEE Access*, vol. 10, pp. 1–20, 2022.

[5] Yamagishi, J. et al., "Thousands of Voices for HMM-Based Speech Synthesis: Analysis and Application of TTS Systems Built on Various ASR Corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 984–1004, 2010.

[6] Kain, A., and Macon, M. W., "Spectral Voice Conversion for Text-to-Speech Synthesis," in *Proc. IEEE ICASSP*, 1998, pp. 285–288.