CrossOracle

An Agentic Framework for Real-Time Expert Personas with Parallelized Acoustic Synthesis

Yash Raj Singh

Independent Researcher

SIGDIAL 2026 System Demonstration

The Problem Space

The Latency Wall in Cascaded Pipelines

Addressing the strict 200ms human turn-taking barrier in Hinglish spoken dialogue systems.

Challenges in Real-Time Dialogue

- **Turn-Taking Constraints:** Natural human conversation demands inter-turn gaps of ~200ms. Cascaded pipelines (ASR → LLM → TTS) struggle to meet this due to additive latency.
- **Hinglish Complexity:** Fluid interleaving of Hindi morphology (SOV) with English loanwords (SVO) breaks standard end-to-end models at scale.
- **Compute Overhead:** Native E2E multimodal models remain computationally prohibitive, lacking the modularity and deterministic control of cascaded architectures.
- **The Goal:** Achieve human-parity latency in a highly modular cascaded system without the "walkie-talkie" delay.

Architectural Innovations

Semantic Boundary Detection

A language-aware heuristic evaluating the token stream for Hindi morphology and Latin punctuation to trigger synthesis mid-thought.

Asynchronous TTS Worker

A First-In-First-Stream (FIFS) queue that decouples and parallelizes ElevenLabs synthesis with ongoing LLM generation.

Deterministic Barge-in

A synchronous hardware-level abort protocol that clears audio buffers instantly, bypassing probabilistic VAD delays.

Semantic Boundary Detection Heuristic

The checkFlush() Logic

Evaluated entirely in the token callback with $\mathcal{O}(1)$ regex operations. Triggers on three classes:

- **Hindi Morphology:** Devanagari *danda* ($।$) or double *danda* ($॥$).
- **Latin Punctuation:** Sentence-final marks (?, !) bypassing abbreviation blocklists.
- **Structural Boundaries:** Em-dashes and explicit newlines.

Preserving Prosody

Streaming TTS (like FastSpeech 2) generates partial waveforms, destroying prosody without future context.

CrossOracle bridges this by flushing at clause-level boundaries, delivering complete prosodic units while aggressively reducing Time-to-First-Byte (TTFB).

Deterministic Barge-in Protocol

< 1ms

ABORT LATENCY

Eliminating "Ghost Audio"

User interruptions trigger a WebSocket message that invokes a global AbortController.

This synchronously cancels the active LLM streaming Promise, halts in-flight TTS requests, and clears the queue.

The finite state machine returns to Idle instantly, preventing the overlapping audio artefact common in cloud-based SDS pipelines.

ASR Error Tolerance

Hinglish ASR often hallucinates Latin strings. We filter these pre-inference using a ratio formula:

$$r = \frac{c \in [\ U{+}0900 \ldots U{+}097F\ ]}{non\_whitespace\_chars}$$

- Transcripts with $r < 0.1$ and length > 8 are discarded silently.
- Empirical separation in logged turns: Accepted transcripts $r = 0.61$ vs. Hallucinated $r \approx 0.02$.
- LLM is primed with an ASR error-tolerance system prompt for contextual repair on residual noise.

The Expert Persona Engine

Preset Identities

Supports 11 highly specific expert identities (Doctor, Lawyer, Strategist, etc.).

- Distinct ElevenLabs voice identities.
- System prompts enforcing Devanagari script output.
- Edge-cached greetings for zero-latency instantiation (866ms vs 4.9s cold start).

Custom Persona Forge

An open-ended routing engine for on-the-fly persona creation from natural language descriptions.

- Extracts name and voice identity from free-text.
- Live keyword-to-voice-bucket routing.
- Enables novel persona construction during live demos.

Live Demonstration Scenarios

Persona Switching

Observe seamless transitions across 11 preset experts with instantaneous context resets and distinct voice profiles.

Barge-in Stress Test

Interrupt the system mid-response to experience the < 1ms deterministic buffer flush and elimination of ghost audio.

Hinglish Parsing

Test dense intra-sentential Hinglish to observe the Devanagari ratio filter and language-aware semantic chunking in action.

Looking Ahead

Conclusion & Future Work

Proving highly responsive conversational latency is achievable in cascaded Hinglish SDS.

Future focus: Formal Mean Opinion Score (MOS) evaluations across diverse Hinglish demographics.

Questions?

Thank you for your attention.

Access the live system demo at
https://crossoracle.io