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

Yash Raj Singh
Independent Researcher
yashmanuraj2@gmail.com

Abstract

We present CrossOracle, a production-deployed spoken dialogue system (SDS) optimized for real-time Hinglish (Hindi-English code-switched) voice interaction across eleven distinct Expert Persona identities. The system addresses the Latency Wall inherent to cascaded Automatic Speech Recognition (ASR) $\rightarrow$ Large Language Model (LLM) $\rightarrow$ Text-to-Speech (TTS) pipelines by introducing a parallelized asynchronous TTS queue and a language-aware semantic chunking heuristic. Key empirical results across 8,000 conversational turns include a mean TTS Real-Time Factor (RTF) of 0.107, a mean Time-to-First-Byte (TTFB) of 4.9s, and a barge-in resolution rate of 100% with a mean abort latency of < 1 ms. The live system is accessible at https://crossoracle.io.

1 Introduction

This paper accompanies a live system demonstration at SIGDIAL 2026 and describes the architecture enabling real-time Hinglish spoken dialogue at the conference booth. Human turn-taking norms impose strict timing constraints on spoken dialogue systems: natural inter-turn gaps average approximately 200 ms (Levinson, 2015).[1] While end-to-end Speech-to-Speech (S2S) models such as AudioPaLM (Rubenstein et al., 2023) attempt to bypass text-generation bottlenecks, they remain computationally prohibitive at scale and struggle with dense intra-sentential code-switching (Sitaram et al., 2019). While recent native E2E multimodal models offer theoretical latency advantages, cascaded pipelines retain practical advantages in modularity, domain adaptability, and deterministic control over persona guardrails without requiring massive, specialized compute infrastructure.

This challenge is acute for Hinglish, where speakers fluidly interleave Hindi morphology (Subject-Object-Verb (SOV) order) with English loanwords (Subject-Verb-Object (SVO) order) within a single utterance. Cascaded pipelines—ASR followed by an LLM followed by TTS—retain practical advantages in modularity and domain adaptability, but their additive latency renders them ill-suited for natural conversation without architectural mitigation. CrossOracle directly addresses this gap through a First-In-First-Stream (FIFS) worker that parallelises TTS synthesis with ongoing LLM generation, a Hinglish-aware semantic boundary heuristic, and a deterministic interrupt protocol.

The demonstration invites conference attendees to hold unscripted Hinglish conversations with eleven expert identities—spanning medicine, law, finance, academic mentoring, and companionship personas—through a browser-based interface, using noise-cancelling headsets provided on-site.

2 Related Work

Streaming TTS architectures such as FastSpeech 2 (Ren et al., 2021) generate partial waveforms before full text is available, but naive word-level streaming destroys prosody by depriving the synthesis engine of future intonation context. Sentence-level buffering preserves prosody at the cost of high latency. CrossOracle bridges this trade-off through clause-level semantic flushing that delivers complete prosodic units while aggressively reducing TTFB.

On Hinglish ASR, prior work has explored language identification and acoustic model adaptation for code-switched speech (Sitaram et al., 2019), but production-grade dialogue systems operating on Hinglish in real time—without a native end-to-end S2S model—remain underexplored. CrossOracle contributes an empirically validated, deployment-

[1] Cross-linguistic studies report median inter-turn gaps of 0–200 ms across typologically diverse languages, suggesting this norm is broadly universal (Stivers et al., 2009).

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 254–257
August 2–5, 2026.

ready systems-integration architecture for this setting, extending recent parallelization efforts like PSLM (Mitsui et al., 2024) into production constraints. Recent work on streaming dialogue includes Radford et al. (2023) for robust multilingual ASR and **?** for low-latency neural TTS. The growing deployment of voice agents (OpenAI, 2023) further motivates latency-optimized cascaded architectures.

3 System Architecture

CrossOracle operates as a Node.js WebSocket middleware bridging Azure Cognitive Services (ASR, LLM) and ElevenLabs Turbo v2.5 (TTS). Audio is transmitted as 16 kHz Pulse-Code Modulation (PCM) over bidirectional WebSockets. The pipeline has three novel components described below.

3.1 Hinglish Semantic Boundary Detection

The `checkFlush()` heuristic determines when to dispatch the current text buffer to the TTS engine. It evaluates the incoming LLM token stream for three trigger classes, in priority order:

1. **Hindi morphology.** Detection of the Devanagari danda () or double danda (), marking a hard sentence boundary in Hindi script (Kaur et al., 2020).

2. **Latin punctuation.** Sentence-final `.`, `?`, or `!` appended to strings of three or more alphabetic characters. A curated abbreviation blocklist (Dr., Mr., etc.) prevents premature flushing.

3. **Structural boundaries.** Explicit newlines (`\n`), em-dashes, and ellipses, which commonly delimit clause boundaries in Hinglish dialogue.

A minimum buffer length of 30 characters prevents sub-utterance chunks from reaching the TTS engine. A hard cap at 220 characters ensures bounded latency regardless of punctuation density. This heuristic is evaluated entirely in the token callback with $\mathcal{O}(1)$ regex operations, adding zero measurable overhead to the streaming loop.

3.2 Asynchronous TTS Worker

As the LLM streams tokens, flushed semantic chunks are enqueued into a `ttsQueue`. A concurrent background worker dequeues each chunk and dispatches it to the ElevenLabs API independently of the LLM generation loop. The resulting 16 kHz PCM binary is forwarded to the client immediately upon receipt. The perceived latency collapses to:

$$L_{perceived} = L_{ASR} + T_{TTFT} + T_{chunk} + L_{TTS}(chunk_1) \tag{1}$$

where T_{TTFT} is the LLM time-to-first-token and T_{chunk} is the time to accumulate the first semantic phrase. Subsequent chunks are synthesised concurrently with playback of earlier chunks. With a mean RTF of 0.107, synthesis completes nearly an order of magnitude faster than real-time playback, guaranteeing the continuous playback condition:

$$\sum_{i=1}^{k} T_{synthesis}(i) < \sum_{i=1}^{k-1} D_{audio}(i) \tag{2}$$

This was satisfied across all 8,000 turns in the evaluation window, with zero instances of mid-sentence acoustic buffering.

3.3 Deterministic Barge-in Protocol

User interruptions are signalled via a WebSocket `interrupt` message, triggered either by Voice Activity Detection or a manual User Interface (UI) button. Upon receipt, the server invokes a global `AbortController`, which synchronously cancels both the active LLM streaming Promise and any in-flight ElevenLabs requests, and clears the `ttsQueue`. The finite state machine returns to Idle in under 1 ms (measured as internal server-side execution time of the `AbortController`), preventing the "ghost audio" artefact common in cloud-based SDS pipelines.

3.4 Hinglish ASR Error Tolerance and Mishear Filter

The Azure Speech-to-Text (STT) engine is statically locked to the `hi-IN` locale, bypassing language identification latency. To handle hallucinated Latin strings—a known failure mode of Hinglish ASR—each transcript is passed through a Devanagari ratio filter:

$$r = \frac{|\{c \in [U+0900 \ldots U+097F]\}|}{non - whitespacechars} \tag{3}$$

Transcripts with $r < 0.1$ and length > 8 characters are discarded silently, and a `stt_mishear` signal is emitted to the client. In the evaluation

corpus, accepted transcripts exhibited a mean r of 0.61, providing a clear empirical separation from the hallucinated-Latin distribution ($r \approx 0.02$). The LLM is additionally primed with an ASR error-tolerance system prompt instructing it to perform phonetic contextual repair on residual transcription noise.

3.5 Expert Persona Engine

CrossOracle supports eleven preset Expert Personas (Doctor, Lawyer, Therapist, Career Coach, Astrologer, Monk, Strategist, Entrepreneur, Wealth Manager, Debate Partner, Drill Sergeant) and an open-ended custom persona engine. Each persona carries a distinct ElevenLabs voice identity, a persona-specific system prompt enforcing Devanagari script output for Hindi lexemes, and a pre-synthesised greeting cached at the edge for zero-latency persona instantiation (mean greeting latency: 866 ms, compared to 4,900 ms for a cold LLM turn). Persona switching is instantaneous from the user's perspective.

The custom persona engine extracts a persona name and voice identity from a free-text description using a keyword-to-voice-bucket routing table, enabling attendees to construct novel personas on-the-fly during the demonstration.

4 Evaluation

Table 1 summarises pipeline performance aggregated across 8,000 multi-domain conversational turns spanning all preset personas active at evaluation time. Telemetry was captured via millisecond-resolution server-side session logs recording per-chunk synthesis times, audio durations, flush trigger reasons, and interrupt resolution latencies.

Metric	Naive Buffer	CrossOracle
Mean TTFB (s)	6.5	4.9
Mean TTS RTF	0.134	0.107
Barge-in Rate	-	100%
Abort Latency (ms)	-	<1
Greeting Latency (ms)	-	866

Table 1: Pipeline performance across 8,000 logged turns. Naive Buffer refers to fixed-token-count flushing with no semantic boundary detection. $RTF = T_{synthesis}/D_{audio}$: values < 1 guarantee gapless playback.

A notable edge case was observed when the LLM generated an isolated Unicode emoji as a final token. The TTS engine required 233 ms to produce 93 ms of audio ($RTF = 2.513$), briefly inverting the playback guarantee at end-of-turn. To address this, we have since deployed a pre-TTS regex filter in production to strip non-lexical Unicode tokens before synthesis, ensuring the continuous playback condition holds across all edge cases.

While this demonstration paper focuses on quantitative systems telemetry (RTF, TTFB), formal subjective evaluations of prosodic naturalness (Mean Opinion Score, or MOS) are planned for future work.

5 Demonstration

Setup. Attendees interact with CrossOracle via a browser-based interface at `https://crossoracle.io` using noise-cancelling headsets provided at the demonstration booth. No installation is required. To ensure seamless accessibility and frictionless onboarding for all conference attendees, the platform natively integrates Google OAuth. Alternatively, reviewers may use the following test credentials: `sigdial_reviewer@crossoracle.io` (Password: `Demo2026!`).

To accommodate the U.S.-based conference demographic, the system's underlying instruction-following architecture natively supports pure English interactions. Attendees can simply instruct the active Expert Persona to "speak in English," demonstrating the LLM's ability to dynamically override its default Hinglish guardrails in real-time while maintaining the continuous playback guarantee.

Scenario 1 – Expert Persona Switching. Participants select from eleven Expert Personas and engage in free-form queries (e.g., medical, legal, financial). The pre-synthesised greeting plays within 866 ms. Participants are encouraged to instruct the persona to switch languages dynamically or switch to a different persona mid-session to observe instantaneous context resets and voice identity changes.

Scenario 2 – Barge-in Stress Test. Participants interrupt the system mid-response using the UI button or by speaking over the AI. The deterministic `AbortController` clears the audio pipeline in under 1 ms, eliminating ghost audio. This is the most viscerally noticeable feature for attendees accustomed to conventional SDS pipelines.

Scenario 3 – Custom Persona Forge. Participants describe an arbitrary persona in natural language (e.g., "a sarcastic Mughal emperor who gives stock advice"). The system performs live

keyword-to-voice-bucket routing, assigns an appropriate ElevenLabs voice, and generates a contextually appropriate greeting within the standard cold-start latency budget.

Scenario 4 – Hinglish Code-Switching. Participants speak in dense intra-sentential Hinglish (e.g., mixing "Mera portfolio diversify karna chahiye" with English domain terms). The semantic chunking heuristic preserves prosodic integrity across script-switching boundaries, which participants can observe in the real-time transcript displayed on-screen.

6 Conclusion

CrossOracle demonstrates that highly responsive, production-viable conversational latency is achievable in cascaded Hinglish SDS without end-to-end Speech-to-Speech models. The combination of FIFS parallelism, language-aware semantic chunking, and deterministic barge-in management produces a qualitatively fluid experience across eleven persona identities. The system is production-deployed and available for live interaction. Recent production updates have introduced persistent cross-session memory for Expert Personas via vector database integration, and future work will focus on a formal MOS evaluation across diverse Hinglish speaker demographics.

References

Jaspreet Kaur, Gurpreet Singh, and Rajneesh Sharma. 2020. A review of Devanagari script and its challenges. *Journal of Indic Languages*.

Stephen C Levinson. 2015. Turn-taking in human communication: Origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Kentaro Mitsui, Soumi Maiti, and Shinji Watanabe. 2024. PSLM: Parallel generation of text and speech with LLMs for low-latency spoken dialogue systems. *arXiv preprint arXiv:2406.12428*.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Wolfgang Macherey, David Marsh, Daisy McGuffey, Tim McVeigh, Warren Morningstar, Ryosuke Nakamura, Khe Pipatsrisawat, Andrew Rosenberg, Arun Rzayev, and 5 others. 2023. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sravana Kumar Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, and Stephen C Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.