Interruption

Voice Agent Networking Event

Definition

Interruption in voice agents occurs when users speak while agents are speaking, or when agents cut off users mid-sentence. Natural human conversation includes cooperative interruptions for clarification, corrections, and turn-taking.

Types of Interruption

Barge-in allows users to interrupt agents. Essential for correcting errors or providing urgent input. Without barge-in, users must wait through irrelevant responses.

Agent interruption happens when VAD triggers too early, cutting users off mid-utterance. Causes frustration and requires users to repeat themselves. Often results from aggressive VAD thresholds.

Cooperative interruption occurs naturally when both parties speak simultaneously to clarify or confirm. Humans handle this smoothly through social cues. Voice agents struggle without visual feedback.

Barge-In Implementation

Full barge-in stops agent speech immediately when user starts speaking. Provides maximum responsiveness but may discard useful information. Best for transactional interactions where speed matters.

Soft barge-in continues agent speech at reduced volume while monitoring user intent. If user persists, full interruption triggers. Balances responsiveness with information delivery.

Context-aware barge-in uses conversation state to determine interruption handling. During critical information delivery, higher thresholds prevent accidental interruption. During small talk, lower thresholds enable natural flow.

Detection Mechanisms

Audio-level detection triggers on sound above noise threshold. Simple but prone to false positives from background noise. Requires 200-300ms of speech to confirm genuine interruption.

Voice activity detection improves accuracy by distinguishing speech from other sounds. ML models identify human speech patterns, reducing false triggers from coughs, ambient noise, or other speakers.

Intent-based detection analyzes partial speech to determine interruption purpose. Quick affirmations (okay, yes, mm-hmm) may not require full interruption. Corrections or questions trigger immediate response.

Interruption Handling

Immediate stop halts agent speech and begins processing user input. Audio buffers flush instantly. Provides fastest response but may feel abrupt. Best for corrections and urgent requests.

Graceful completion finishes current sentence before stopping. Adds 500-1500ms delay but feels more natural. Appropriate when agent is delivering critical information or making key points.

Buffered processing captures user speech while agent continues, then processes after agent finishes. Allows non-urgent interjections without disrupting flow. Risky if user expects immediate response.

False Positive Management

Confirmation windows require 150-300ms of sustained speech before triggering interruption. Prevents brief sounds or crosstalk from stopping agents unnecessarily.

Energy thresholds distinguish speech from background noise. Adaptive thresholds adjust to ambient conditions. Quiet environments use lower thresholds, noisy environments require higher energy to trigger.

Speaker diarization identifies if interrupting speech comes from original user versus other speakers. Prevents conversations in background from triggering false interruptions. Requires consistent voice fingerprinting.

User Experience Considerations

Clear feedback indicates when interruption succeeds. Audio beeps, visual cues, or verbal acknowledgment confirms agent is listening. Without feedback, users don't know if interruption worked.

Recovery mechanisms handle failed interruptions gracefully. If system misses interruption, provide clear way to retry. Fallback to explicit stop commands (stop, cancel, nevermind) ensures users maintain control.

Context preservation maintains conversation state across interruptions. When users interrupt to ask related questions, agents should remember original topic and offer to continue.

Technical Challenges

Full-duplex audio requires simultaneous capture and playback. Echo cancellation prevents agent speech from triggering false interruptions. Acoustic echo cancellation (AEC) is critical but adds 50-100ms processing latency.

Latency compounds interruption challenges. Network delays mean 100-300ms passes between user speech and agent response. During this window, agent continues speaking, creating awkward overlap.

Streaming synthesis complicates interruption. TTS may have already generated 500-1000ms of audio buffered for playback. Flushing buffers requires coordination across TTS, audio pipeline, and network layers.

Advanced Techniques

Predictive interruption anticipates user intent to interrupt based on prosodic cues. Rising intonation or increased speaking rate signal impending interruption. System can pre-emptively prepare to stop.

Multi-modal detection combines audio with visual cues when available. Video enables detecting raised hands, leaning forward, or mouth movements before speech begins. Adds 200-500ms advance warning.

Sentiment analysis identifies frustrated interruptions versus engaged questions. Frustrated users receive acknowledgment and modified responses. Engaged users get direct answers.

Design Patterns

Progressive disclosure delivers information in chunks with natural pause points. Each chunk completion creates interruption opportunity. Reduces need for barge-in while maintaining engagement.

Explicit turn-taking uses verbal cues indicating agent finished. Phrases like 'What do you think?' or 'Any questions?' clearly signal user's turn. Reduces ambiguity about speaking rights.

Adaptive verbosity adjusts response length based on interruption history. Users who frequently interrupt get shorter initial responses. Users who listen receive fuller explanations.

Metrics and Evaluation

Interruption rate tracks how often users interrupt. High rates may indicate overly long responses or poor relevance. Low rates might suggest users can't interrupt or aren't engaged.

False positive rate measures incorrect interruption triggers. Target below 5% for acceptable UX. High false positives frustrate users by stopping relevant information.

Recovery time measures how quickly conversation resumes after interruption. Should be under 500ms. Longer delays make interruptions feel costly, discouraging natural interaction.

Best Practices

Enable barge-in by default but make it configurable. Some use cases (emergency instructions, legal disclosures) require uninterruptible speech. Most conversations benefit from interruption capability.

Tune VAD thresholds for specific environments and use cases. Contact centers need different settings than in-car assistants. A/B test threshold values to optimize for your users.

Provide visual feedback when video is available. Show listening indicators, processing states, and confirmation of interruption. Visual cues compensate for audio delays and ambiguity.

Test extensively with real users in target environments. Lab testing misses noise, echo, and behavior patterns. Deploy gradually with monitoring to catch edge cases.