When deploying AI voice agents in contact centers, the perennial question emerges: should you prioritize a solidly clean handoff at around 70% resolution rate, or push harder for an ambitious 80% resolution rate that might increase customer friction? This isn't just an academic debate — it shapes vendor selection, system design, and ultimately affects customer trust and operational efficiency.

Let's carefully unpack this challenge, focusing on key enablers like the telephony stack, speech recognition (Automatic Speech Recognition or ASR), and practical design elements such as end-to-end latency and barge-in (interruption) handling. We'll also explore why legacy IVRs failed us and how the interplay between voice and chat channels creates unique constraints.

Understanding Resolution Rate vs. Handoff Quality

Resolution Rate is the percentage of customer calls that the AI voice agent completes successfully without human intervention. **Handoff Quality** measures how seamlessly the AI transfers the caller to a live agent when needed, including how much re-collecting of information the customer undergoes and whether context is preserved.

The tradeoff: aggressively extending AI coverage to maximize resolution can lead to degraded handoff quality. Conversely, conservative AI use creates more handoffs but each handoff is cleaner and builds customer trust more effectively. So what's optimal?

Why Not Max Out Resolution at 80%?

- **Increased Failure Modes:** Pushing the AI to handle more variants and edge cases raises the risk of errors, misrecognitions, or unsatisfactory resolutions.
- **Higher Latency and Customer Frustration:** Complex dialogues escalate the average handling time—the crucial metric is *end-to-end latency*, not just model reaction time. More back-and-forth prolongs interactions and tests patience.
- **Poor Handoff Experience:** If the AI isn't confident but still tries, handoffs become clumsy and customers have to repeat their issue, destroying trust.

By contrast, resolving 70% with a robust, clean handoff for the remaining 30% can sustain a better overall experience. But you need the right technical foundation.

Legacy IVR Failure: What Went Wrong?

Traditional IVRs typically failed because they:

1. Used **rigid DTMF menus**, limiting natural language input and frustrating callers.
2. Had slow and unresponsive interfaces due to legacy telephony stack constraints.
3. Offered poor or no barge-in functionality, forcing callers to endure full prompts even when they already knew what to say.
4. Produced poor handoff experiences, losing context and forcing repetition.
5. Focused solely on containment rate (resolution) without balancing handoff quality or trust.

These limitations resulted in widespread caller frustration, long average handle times, and high drop rates.

Voice vs. Chat: Different Constraints Matter

Comparing voice to chat helps telephony integration clarify why resolution and handoff dynamics differ.

AspectVoiceChat Input ModalitySpeech, immediate, transientText, persistent, allows review Latency ToleranceNeeds **low end-to-end latency** & quick responsesHigher tolerance to lag; users can read and respond asynchronously Barge-In HandlingEssential to accept interruption mid-promptNot applicable Error HandlingMore sensitive to misrecognition due to speech ambiguityEasier to clarify and retry Handoff ExperienceMust preserve context to avoid repetitionEasier to preserve chat history

Voice AI systems operate under tighter real-time constraints and must skillfully handle interruption ("barge-in") and latency, or customers lose patience quickly. Chat bots can tolerate higher latency with less immediate user frustration.

Telephony Stack and Speech Recognition: Foundations for Success

Telephony Stack Considerations

Your telephony infrastructure profoundly impacts voice AI performance and handoff quality:

- **IP-based Telephony:** Enables easier integration with AI and real-time media streaming.
- **Low Latency Media Path:** Critical to minimize delay from speaking to recognition to response.
- **Barge-In Capable:** Must support mid-prompt interruption without cutting off recognition or crashing the call.
- **Context Aware Routing:** Allow transferring full session state and partial ASR transcripts to human agents.

Speech Recognition (ASR) Quality

ASR accuracy is a linchpin for resolution success. Key notes:

- Models must be tuned for your domain and dialect diversity.
- Noise-robust front-end processing reduces errors.
- Partial or incremental hypotheses facilitate quicker responses and true barge-in support.
- Error detection and fallback mechanisms (like slot confirmation or gracefully handing off) enhance trust.

The End-to-End Latency Imperative

Whenever an AI vendor highlights model latency, push back and ask for the **end-to-end latency** number — the total time from when a caller speaks to when they hear an intelligible response. This includes:

1. Audio capture and transmission time.
2. ASR processing, including decoding partial transcripts.
3. Language understanding and dialogue management.
4. TTS (Text-to-Speech) generation and playback on caller side.

If end-to-end latency creeps above 500-700 ms, caller frustration rises sharply. This impacts how high you can push resolution rates or how elegantly handoffs occur without irritating delays.

Barge-In and Interruption Handling: A Non-Negotiable

One frequent source of failure in voice AI systems is poorly implemented barge-in: the caller's ability to interrupt the system prompt instantly to speed resolution or correct errors. Vendors who dodge questions about barge-in are suspect.

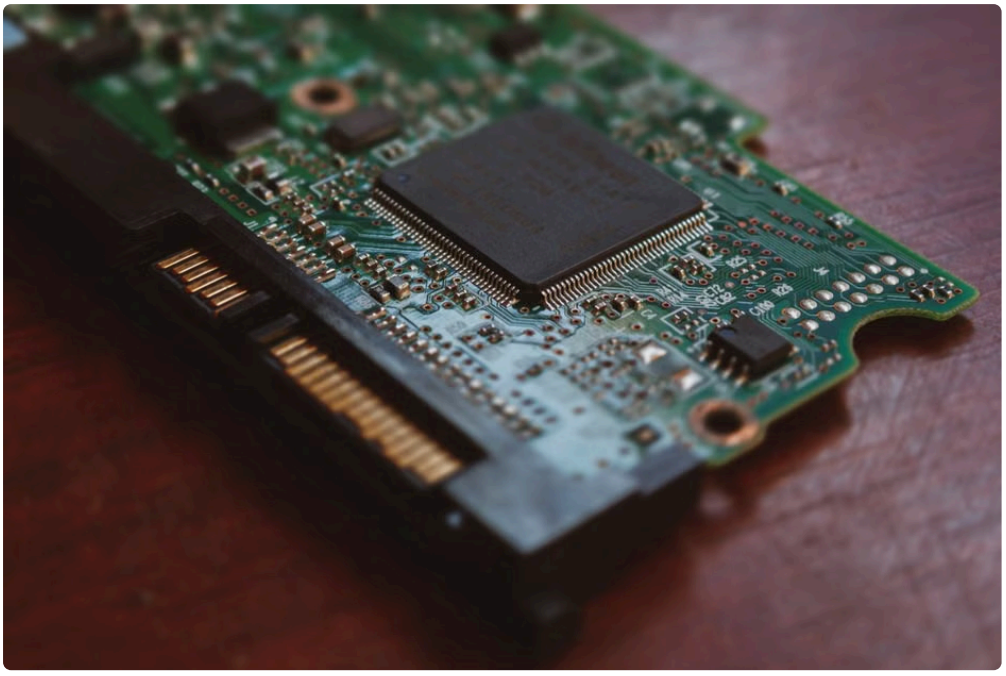
Good barge-in implementation means:

- Real-time detection of speech overlapping system prompts.
- Seamless transition from playing audio to active ASR listening.
- Preserving partial inputs and avoiding premature cut-offs.
- Reducing overall dialogue latency and user frustration.

Systems lacking this will force users to wait unnecessarily, damaging trust and inflating handle time.

Balancing 70% Resolution With Clean Handoff vs. Pushing 80%

From years of piloting voice AI solutions in mid-market retail and healthcare, here's my pragmatic take:



ApproachProsCons 70% Resolution + Clean Handoff

- Higher customer trust due to smooth, context-rich handoffs
- Lower risk of failure modes and frustration
- Better control over average handle time and latency
- Faster ROI via stability and quality metrics
- More handoffs and potentially higher human agent load
- Requires excellent handoff integration and telephony stack

80% Resolution (Aggressive AI Handling)

- Fewer handoffs, increased automation containment metrics
- Potential cost savings if implementation is flawless
- Higher end-to-end latency reduces customer satisfaction
- Increased failure modes and longer dialogues

- Poor handoff quality when handoff happens — mistrust
- Risk of optimizing for numbers rather than real experience

Driving Trust: The Ultimate KPI

Resolution rate and handoff quality must ultimately serve **customer trust.** If callers feel the system "gets it" and treats them respectfully, they're more likely to engage fully and accept automation. Conversely, robotic AI that forces customers to repeat or wait erodes trust and drives them to abandon calls or channels.

Trust is achieved by:

- Designing for clean, informative, context-rich handoffs.
- Ensuring low end-to-end latency and responsive interaction.
- Supporting smart barge-in and natural interruption handling.
- Setting realistic resolution targets aligned with technology capabilities.

Testing for Failure Modes: A Recommended Short List

In every pilot, I insist on testing at least these failure scenarios:

1. Caller interrupts mid-prompt (barge-in test).
2. Recognition failure with noisy background.
3. Partial resolution with immediate human handoff.
4. Context preservation across handoff (no repetition).
5. High latency or network jitter effects on voice interaction.

Testing these scenarios reveals real-world robustness beyond marketing claims.

Final Thoughts

In the balance between resolution rate and handoff quality, experience strongly supports prioritizing a *robust 70% resolution rate with clean, trust-building handoffs* over pushing for 80% automation that risks frustrating customers. Success depends on a solid telephony stack that supports low latency and barge-in, ASR tuned for your voice environment, and a design philosophy that respects user patience and trust.

Your technology investments and vendor selections should reflect these realistic priorities rather than chasing inflated resolution numbers that mask a poor customer experience.

Remember: **It's better to resolve less but do so well than to promise more and deliver frustration.**