If your voice agent keeps talking while a customer is trying to speak, the problem usually isn't that the customer is "interrupting." The agent is failing to detect, interpret, or respond to a change in who has the floor. That can happen because of slow speech recognition, poorly tuned barge-in handling, long answers, or a knowledge system that makes the agent hesitate.

In nine years of CX systems work, I've learned not to trust a pristine demo. I test the same phone lines twice: once in a quiet room and once with background noise. I also keep a list of awkward caller scripts—people who change their minds mid-sentence, say "wait," or start answering before the agent has finished. Those calls reveal whether an agent can actually share a conversation.

This guide walks through the voice agent pipeline, common failure points, and practical ways to improve interruptions settings, turn-taking tuning, and the overall caller experience.

First, identify where the interruption goes wrong

A voice agent is a chain of systems, not a single talking model. A simplified call looks like this:

1. **Audio capture:** The phone line or microphone sends the caller's audio to the system.
2. **Voice activity detection:** The system estimates whether someone is speaking or whether there is silence.
3. **Speech recognition:** Audio is converted into words, often before the caller has finished speaking.
4. **Conversation logic:** The agent decides whether to answer, ask a question, look something up, or escalate.
5. **Knowledge retrieval:** If needed, the system searches a knowledge base or other data source.
6. **Response generation:** The agent prepares the words it intends to say.
7. **Speech generation and playback:** Text becomes audio and is sent back to the caller.

Talking over someone can begin at any step. For example, voice activity detection may mistake a pause in the caller's sentence for the end of their turn. Or the agent may receive the caller's words correctly, but playback continues after the caller says, "Actually, never mind." The fix depends on where the handoff fails.



What you hear Likely failure point What to inspect The agent answers before the caller finishes a thought End-of-turn detection or response timing Silence thresholds, endpointing, partial transcript behavior The caller speaks, but the agent's audio keeps playing Barge-in handling or playback control Interruption detection, playback cancellation, audio buffering The agent stops, then responds to the wrong thing Transcription or turn-state handling Transcript timestamps, dropped words, conversation state The agent pauses for a long time before speaking again Latency or retrieval Speech recognition, model response time, knowledge search, speech generation

Understand latency and realistic turn-taking

People don't wait for a formal "your turn" signal in everyday conversation. They use pauses, emphasis, and context. A voice agent has to estimate when the aiagentstore caller is done, then prepare and deliver a response quickly enough to feel attentive—but not so quickly that it cuts off a thought.

That balance is harder than it sounds. If the agent waits too long after every pause, callers experience awkward silence. If it responds too quickly, it may treat "I need to change my..." as a complete request. Streaming speech recognition can help by providing partial results while the caller is talking, but acting on partial results too aggressively can cause premature replies.

Measure the full path, not just the model's response time. Record timestamps for the end of the caller's speech, the moment the system decides the turn is complete, the start of generated audio, and the moment playback stops. I count dead-air seconds out loud during evaluations because a delay that looks small in a dashboard can feel enormous on a phone call.

Use the right controls for the right problem

Platform names and settings vary, but look for controls related to:

- **Endpointing:** How much silence the system requires before deciding the caller has finished.
- **Speech thresholds:** How much audio activity counts as a person speaking, including quieter speech.
- **Response delay:** Whether the agent waits briefly to see if the caller continues.
- **Playback cancellation:** Whether generated speech stops immediately when the caller begins speaking.
- **Turn state:** Whether the system can distinguish listening, processing, speaking, and interrupted states.

Don't tune these in isolation. A longer endpointing delay might prevent premature replies but make the agent feel sluggish. A very sensitive speech threshold might detect a quiet interruption, but also react to room noise. Make one change at a time, then replay the same test calls.

Make barge-in handling actually work

Barge-in handling is the system's ability to let a caller speak while the agent is speaking—and to respond appropriately. It is more than detecting that sound has started. The system needs to stop or pause the agent's playback, listen to the caller, capture the new words, and decide whether to continue, revise, or abandon its previous response.

A common failure is "detect but don't cancel": the system notices the caller, yet buffered audio keeps playing for another second or two. Another is "cancel but don't understand": playback stops, but the caller's interruption is clipped or ignored. Both are frustrating because the customer has done the work of getting the agent's attention without gaining control of the conversation.

Practical interruptions settings to test

- **Can the caller interrupt every response?** Consider allowing interruption during most spoken answers, with clear exceptions for short confirmations or safety-critical messages.
- **How quickly does playback stop?** Check for buffering between the voice service and the phone connection. Test the actual phone lines, not only a browser preview.
- **What happens to the unfinished answer?** The agent should not resume a stale response after addressing the new question.
- **Does the interruption have to contain words?** A cough, background conversation, or line noise should not necessarily cancel playback. Test sensitivity with real environments.
- **Can the agent acknowledge a correction?** If a caller says, “No, the other account,” the next response should reflect that correction—not repeat the original answer.

For example, if the agent is explaining a return policy and the caller says, “Wait, I bought it in store,” the agent should stop, register the correction, and answer for an in-store purchase. It should not finish a paragraph about mail-in returns first.

Improve turn-taking with shorter, clearer responses

Even strong interruption detection can’t compensate for an agent that speaks for 30 seconds at a time. Long monologues create more opportunities for callers to jump in, and they make it harder to notice when the customer wants to redirect the conversation.

Give the agent a conversational structure: answer the immediate question, pause, and invite the next step. Prefer “Yes, your refund was issued yesterday. Would you like me to check when it should arrive?” over a lengthy explanation of the entire refund process.

Write dialogue rules, not just personality instructions

“Be friendly and natural” is too vague to control turn-taking. Add specific behavior such as:

- Ask one question at a time.
- Keep routine answers to one or two spoken sentences before checking whether the caller needs more.
- When the caller starts speaking, stop playback and listen rather than trying to finish the sentence.
- Treat “wait,” “no,” “actually,” and “let me explain” as likely signals to yield the floor.
- If you are unsure whether the caller has finished, ask a brief clarifying question instead of guessing.

These instructions help, but they are not a substitute for audio-level interruption controls. An agent can be told to yield and still talk over a customer if its playback does not stop.

Check whether the knowledge base is adding delay or confusion

Knowledge bases and retrieval-augmented generation (RAG) can improve accuracy, but they also add steps between the caller’s question and the agent’s answer. A slow search can create dead air. Poorly chunked or conflicting documents can make the agent retrieve the wrong policy, then produce a long, hesitant response. That can sound like a turn-taking problem when the real issue is retrieval quality.

Review the sources the agent retrieves for actual calls. Check whether they are current, specific, and consistent. Break long documents into sections that each answer a clear question. Keep policies,

exceptions, and effective dates easy to distinguish. If several documents disagree, resolve the conflict at the source rather than hoping the model will infer the right answer.

Also give the agent a useful way to handle a slow lookup. It can briefly say, “I’m checking that now,” then return with the answer. But avoid repeated filler phrases or silence with no explanation. For urgent or account-specific requests, make sure the system can use approved integrations or transfer the customer with relevant context.



Test the caller experience—not just the happy path

Run structured tests on both phone lines and your website widget. A website widget is useful for checking the conversation design and text or audio behavior in a controlled environment, but it does not reproduce every phone condition. Phone calls introduce network delay, audio compression, speakerphone echo, and background noise.

Use a small test set before changing settings and repeat it after each change. Include ordinary callers and “awkward caller” cases such as:

- A caller who pauses briefly in the middle of a sentence.
- A caller who says “No” or “Wait” while the agent is speaking.
- A caller who corrects a name, date, or account detail.
- A caller speaking quietly, quickly, or over background noise.
- A caller who changes the request after the agent starts answering.
- A caller who asks a question that requires a knowledge-base lookup.

Score each recording on whether the agent stopped promptly, captured the interruption, understood the correction, and continued without restarting an irrelevant answer. Track latency and incorrect turn-end decisions as well. “It sounded fine” is not a reliable test result unless you can say which calls you tested and what passed.

Connect the voice experience to your support workflow

If the agent cannot resolve a request, escalation should preserve the conversation. A handoff that sends a caller to a person without the issue, details already collected, or steps already attempted makes customers

repeat themselves—and can make a voice experience feel broken even when its speech recognition is working.

For a Zendesk workflow, pass a concise summary, the customer's stated goal, relevant identifiers, and any attempted troubleshooting into the ticket or agent handoff. Make clear which details are confirmed and which are uncertain. If the voice platform supports context transfer, test the handoff end to end rather than assuming the integration passes everything.

Tools such as ElevenLabs and ElevenAgents can be part of a voice setup, but the product name alone won't guarantee good turn-taking. Evaluate the controls available in your specific configuration: can you adjust interruptions settings, inspect timing, cancel playback, and test different call conditions? The answer matters more than a polished sample recording.

A practical troubleshooting sequence

1. **Save representative calls.** Include cases where the agent interrupted, ignored an interruption, or left a long pause.
2. **Locate the failure.** Compare audio, transcripts, turn events, retrieval timing, and playback timestamps.
3. **Fix playback cancellation first** if the caller speaks but the agent audio continues.
4. **Tune endpointing** if the agent answers before a caller has finished. Recheck that the added wait does not create excessive silence.
5. **Shorten spoken responses** and add explicit instructions for corrections and floor-yielding.
6. **Audit retrieval** if answers are delayed, inaccurate, or needlessly long after a lookup.
7. **Retest on real calls** with quiet and noisy conditions, then monitor for regressions.

The goal is control, not a "human-like" demo

A good voice agent does not have to sound magically human. It has to be responsive, accurate, and willing to give the caller the floor. That means measuring the full pipeline, tuning barge-in handling and endpointing together, keeping answers concise, and making knowledge retrieval dependable.

Start with recordings, find the exact point where turn control breaks, and change one setting at a time. When callers can interrupt, correct the agent, and get a relevant response without repeating themselves, you'll see the difference in the caller experience—and hear it in the calls.