

Voice AI for customer support: what works

What voice agents genuinely handle in a support operation, and where they should never be pointed.



Amit Kumar

Founder, Udyat Technologies

THE SHORT VERSION

Point Voice AI at the repetitive, verifiable calls. Measure resolution, not deflection. And make the handover to a human feel like help, not failure.

- ✓ The win is in status, scheduling, and verification calls — not in complex support.
- ✓ Containment is a vanity metric if the caller rings back an hour later.
- ✓ The handover to a human decides whether customers accept the system at all.
- ✓ In India, language switching mid-call is the norm and has to be designed for.

Voice AI has moved a long way in two years. The speech recognition is good, the latency is low enough to feel conversational, and the cost per call is a fraction of a human minute. None of that is the hard part any more.

The hard part is scope. Almost every disappointing voice deployment we have been asked to look at failed the same way: it was pointed at the whole support queue instead of the part of it that is genuinely repetitive, and customers learned within a week that the fastest route to a resolution was to say nothing until they were transferred.

What follows is what we have found actually works, and what we tell people not to bother with.

Where voice agents earn their place

The common thread: a clear question, a verifiable answer, and a system that already holds it.

Status and tracking

Where is my order, when is the engineer arriving, has my payment cleared. High volume, low complexity, and the answer already exists in a system. This is the single strongest use case and usually the first one to build.

Scheduling and reminders

Booking, confirming, rescheduling, and chasing no-shows. Outbound reminder calls in particular pay back quickly because they replace work nobody enjoys doing and everybody skips when busy.

Qualification and routing

Establishing what the caller needs and who should handle it, before a human picks up. Even when the agent resolves nothing, arriving at a human with context already gathered removes a minute from every call.

First-line verification

Confirming identity, order references, or policy numbers against a system of record, so the human conversation starts past the tedious part rather than beginning with it.

Where not to point it

Not because the technology cannot attempt these, but because attempting them costs more in trust than it saves in minutes.

- ✓ Complaints and anything where the caller is already angry.
- ✓ Anything involving a judgement call about money, liability, or an exception to policy.
- ✓ Calls where a wrong answer is expensive to unwind — medical, legal, or regulatory.
- ✓ Cancellations and retention conversations — a person should hear those.
- ✓ Technical diagnosis where the next question depends on interpreting a vague answer.
- ✓ Anything you would not be comfortable explaining to the customer afterwards.

Deflection is not the metric. If the caller rings back an hour later, the call was not contained — it was postponed, and it now costs two calls instead of one.

Measure resolution, not containment

Voice AI platforms report containment: the share of calls that ended without a human. It is the number that appears in every vendor deck and it is close to useless on its own, because a caller who gives up and hangs up counts as contained.

The number worth tracking is resolution within 24 hours without a repeat contact. It is harder to measure, because it means joining call records to tickets and to whatever the customer did next. It is also the only figure that tells you whether the system is helping.

A well-scoped deployment on status and scheduling calls typically resolves 55–70% of the calls it is pointed at, without a repeat. A badly scoped one reports 80% containment and a rising complaint rate. We have seen both, and the second is much harder to recover from, because customers stop trusting the phone line.

The handover decides everything

Customers forgive an agent that cannot help. They do not forgive having to start again.

A handover that works

- Triggered early, on the second sign of difficulty
- The human receives a summary and the full transcript
- Nothing already given is asked for a second time
- The caller is told a person is joining, and why
- An explicit route out: say 'agent' at any point
- Out of hours, it takes a message rather than looping

A handover that does not

- Triggered only after repeated failure to understand
- The human starts cold with no context
- Identity and order number requested again
- The transfer is silent and feels like a dropped call
- No way out except hanging up and redialling
- Out of hours, the loop simply repeats

The language reality in India

Most vendor demos are single-language. Real calls here are not. A caller opens in English, switches to Hindi or Punjabi when explaining the actual problem, and drops back into English for the product name and the order reference. A system that handles each language well in isolation can still fail on the switching.

This is testable before you commit. Take fifty recordings of real calls from your own queue, including the ones with background noise from a shop floor or a street, and run them through whatever you are considering. The gap between a demo and your recordings is usually the whole decision.

Telephony is the other quiet variable. A voice agent that performs well over a clean SIP connection can degrade badly over a poor mobile line or a compressed PRI. Test on the lines your customers actually call in on, not on the ones in the vendor's lab.

How we pilot it

Six to eight weeks, one call type, and a decision at the end that can honestly be no.

Pick one call type

The highest-volume, most repetitive one. Usually order or ticket status. Everything else keeps going to humans exactly as it does today, so there is no downside case.

Build against real recordings

The agent is tuned on your own calls, in your own languages, with your own product names — not on a generic template adjusted afterwards.

Run it on a slice

A portion of the traffic, with an immediate route to a human and a hard rule that anything unclear escalates rather than guesses.

Judge it on resolution

Resolution without repeat contact, average handling time for the calls that do reach a human, and complaint rate. If those three do not move the right way, we say so and stop.

The best outcome is not a phone line without people on it. It is a phone line where the people are spending their time on the calls that actually needed a person.