Microsoft

Voice Agents That Ship: Real-Time AI in Dynamics 365 Contact Center

Peter Ruiter · Microsoft MVP · CTO
Microsoft CX, Capgemini

The MVP who forgot his own password

Locked out at 8:55.
Keynote at 9:00.

New laptop, old password, no idea which one.
The only option: call my own helpdesk and wait in the queue.

40%

of helpdesk calls are password resets

$70

per reset in technician time and lost productivity

Sources: Gartner via ZDNet; Infosecurity Magazine

The one idea

Generative talks. Deterministic ships.

The model owns

- The conversation itself
- Interruptions and barge-in
- Accents, rambling, mid-call language switch
- Deciding which tool to call
- Explaining what happens next

Tools and topics own

- Who is calling (voice, employee number, MFA push)
- Whether verification succeeded
- The exact characters of the pass
- How many retries, and when to stop
- What the human gets at handoff

A tool result decides. The model is never allowed to believe a caller who says they are already verified.

Microsoft Ignite 2026 · Lightning Talk

Voice Agents That Ship

Real-Time AI in Dynamics 365 Contact Center · recorded in the Contact Center Intelligence portal, Preview tab

The latency budget



0 to 1 s
Feels like a conversation

1 to 3 s
Feels like a pause. Fill it.

over 3 s
Feels broken. Callers say "hello?"

Hop	Budget	Guardrail
Voice activity detection	750 ms silence (default)	Semantic VAD for callers who think out loud
Model first audio	under 1 s on speech-to-speech	Short turns. Three sentences max by ear
Dataverse lookup	about 1 s	Say one line first, then silence
Graph: create pass	1 to 2 s	Latency message after 500 ms, configured, not prompted
MFA push	as long as the caller takes	One push. No retry. Timeout decides

EU DATA

Speech-to-speech (GPT-Realtime) processes outside the EU Data Boundary. In-region gets the text model (STT, GPT-5-Chat, TTS) and seconds of silence per turn. Budget for it before you promise anything.

Chapter 4

When it breaks

Second call · language switch · the push that never went out

Failure 1: dead air

WHAT HAPPENS

- The tool call takes 3 s and nothing plays
- The caller says "hello?"
- The agent treats "hello" as new input
- Now you have two conversations

THE FIX

- A configured message on every tool call over 2 s
- Played after 500 ms of silence, by the platform, not the model
- A millisecond budget per hop, and someone who owns it
- Semantic VAD so a thinking pause is not a turn

Failure 2: the model decides who you are

WHAT HAPPENS

- The orchestrator re-runs authentication because the caller mentioned their account
- Or the caller says "my manager already approved me"
- And the model, being helpful, agrees

THE FIX

- Name the tool by purpose: "verify caller before password reset", not "authenticate user"
- Only a tool result sets the verified flag. Put that sentence in the instructions
- A condition node in front of every action that needs it
- One push per call. No retry. Timeout decides

Failure 3: escalation drops context

WHAT HAPPENS

- The handoff works
- The rep says "can I have your employee number?"
- The caller has now said it three times to two parties
- This is the moment people decide the AI is useless

THE FIX

- Set the handoff variables before the Escalate node, every time
- Verified how. What failed. What was already done
- Ten seconds of summary, not a transcript
- If you cannot show it at the node, you are not ready to ship

Where this runs

Copilot Studio

Customer Conversation harness

- Primary agent + instructions (speech, chat)
- Tools: Agent Flows, Escalate
- Skills: generative and topic (deterministic)
- Events: consent, speech collection, silence
- Voice settings: DTMF, VAD, latency message

Contact Center Intelligence

Portal and services

- Biometric authentication policy
- Voiceprints stored in Dataverse, not audio
- Anti-spoofing and synthetic speech detection
- Quality and operational insights
- Connected agents from Azure AI Foundry

Contact center

Where the call lands

- Dynamics 365 Contact Center: voice, routing, rep workspace, handoff with context
- Genesys Cloud via connector
- Other CCaaS via universal connector
- Same agent, same voiceprints

Status today

Copilot Studio	Contact Center Intelligence	Contact center
GA · Apr 2026	Preview	Preview · Oct 2026
Real-time voice agents in D365 CC	Customer Conversation harness	Contact Center Intelligence + biometrics
		GA · Nov 2026
		Foundry agents in CCI (roadmap)

Generative talks. Deterministic ships.

- Let the model own the conversation, including the language
- Identity, the pass, retries and the handoff live in tools and topics. A result decides, not a guess
- Budget the silence. Message every hop over two seconds
- Hand the human a summary, not a transcript
- Build the deterministic parts first. They do not change between preview and GA



Agent, topics, instructions, Entra setup
powerpete.com/ignite-voice-agents

Peter Ruiter
powerpete.com · linkedin.com/in/peterruiter · peter.ruiter@capgemini.com