

Voice AI's price war is being fought with expiry dates

Google, ElevenLabs and Microsoft shipped cheaper speech models within nine days, but every headline rate expires between 12 October and 1 January. Even at list prices, the voice layer of an AI agent now costs a few dollars per conversation-hour. That moves the bottleneck on call-centre automation from the price of speech to reliability and customer trust.

AUTHOR



Alicante Tech Vibes *Research Labs*

~8 MIN READ
DATED

ALL FIGURES SOURCED &

\$9 → \$18 GEMINI FLASH TTS PER M TOKENS, 2026→2027 · DATANORTH, SEP 2026	72% ELEVEN V4 LAUNCH DISCOUNT TO 12 OCT · DIGITAL APPLIED, SEP 2026	\$22 MAI-VOICE-2.1 PER M CHARACTERS · MICROSOFT AI, 1 OCT 2026
\$0.54 MAI STREAMING TRANSCRIPTION PER AUDIO HOUR · MICROSOFT AI, OCT 2026	\$21.53 US CUSTOMER SERVICE MEDIAN HOURLY WAGE · BLS, 2025	20% LEADERS REPORTING AI HEADCOUNT CUTS · GARTNER, DEC 2025

I. Three launches, three expiry dates

On 23 September, Google released [Gemini 3.8 Flash TTS and Flash-Lite TTS](#), covering more than 100 languages and over 2,000 ready-made voices. Its audio output is priced at [\\$9 per million tokens through 31 December](#), rising to [\\$18 from 1 January 2027](#). The doubling appears in the pricing table, not the announcement.

Five days later, ElevenLabs shipped [Eleven v4 and v4 Turbo](#), with Turbo reaching audible speech in a median of about 150 milliseconds. [Digital Applied lists](#) v4 at \$0.08 per 1,000 characters and Turbo at \$0.04, both cut 72% until 12 October.

On 1 October, Microsoft AI released [MAI-Transcribe-2-Streaming](#) at \$0.54 per audio hour "through the end of the year", alongside MAI-Voice-2.1 at \$22 per million characters and MAI-Voice-2.1-Flash at \$15.

Read as a price war, this is a war of introductory offers. Every headline number above has an end date inside the next three months.

II. The speech layer became a hyperscaler feature

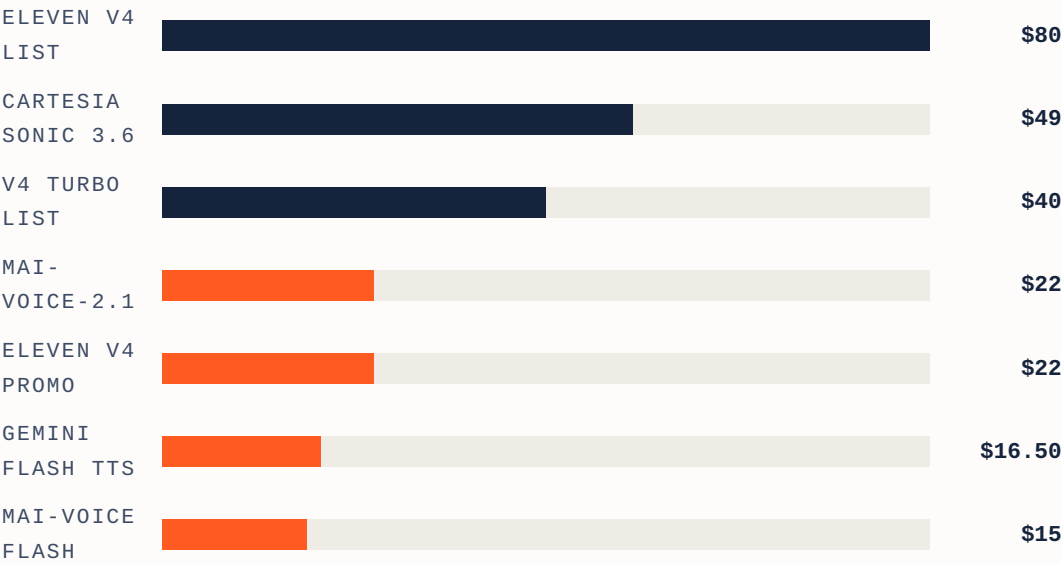
The bigger change is who sells the voice. With the October releases, Microsoft can offer streaming transcription, a language model and speech synthesis inside its own cloud. [Tech Times describes this](#) as completing the voice-agent pipeline, so developers no longer need a third-party transcription vendor. Google already bundles TTS into the Gemini API and named LiveKit, Pipecat, Agora and Vercel as launch partners.

That squeezes specialists in theory. In practice the leading specialist is growing fast. [TechCrunch reports](#) ElevenLabs' annualised revenue rose from about \$330 million at the start of 2026 to more than \$600 million, with 55% from large companies. A hyperscaler entering a market and a specialist doubling revenue in the same year are not contradictory. The market is expanding faster than share is being taken.

Each vendor also claims first place on a different scoreboard. ElevenLabs cites the Artificial Analysis voice arena. Google cites Hume AI's Voice Design Benchmark, where Flash TTS scored 71.4. Microsoft cites Artificial Analysis

for transcription accuracy. These are separate tests of separate things, and no single ranking settles who leads.

FIG. 1: TEXT-TO-SPEECH PRICE PER MILLION CHARACTERS, OCTOBER 2026



BARS SCALED TO THE HIGHEST PRICE (\$80). ORANGE MARKS TEMPORARY OR INTRODUCTORY RATES: ELEVEN V4 PROMO TO 12 OCT, GEMINI TO 31 DEC, MAI INTRO PRICING. GEMINI FIGURE IS ARTIFICIAL ANALYSIS' PER-CHARACTER CONVERSION. SOURCES: DIGITAL APPLIED, 28 SEPTEMBER 2026; MICROSOFT AI, 1 OCTOBER 2026.

III. The honest technical read

The latency gains are real. A full conversational turn on Microsoft's stack takes an estimated 580 to 780 milliseconds, with the language model consuming most of it. Speech is no longer the slow part of a voice agent.

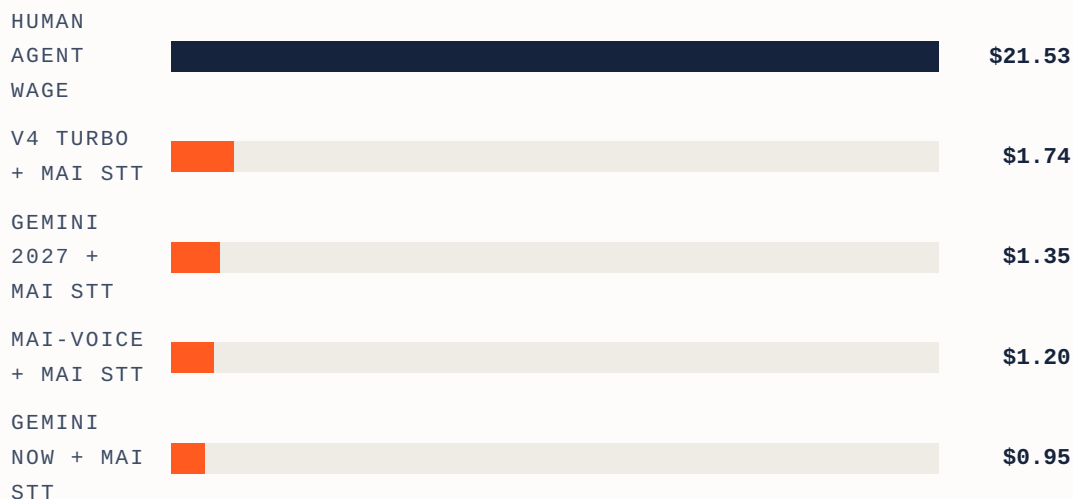
The production case is weaker than the launch posts suggest. Rajesh Beri's 3 October review notes that MAI-Transcribe-2-Streaming is a preview Microsoft does not recommend for production, with no service-level agreement, no word-level timestamps or confidence scores, and preview terms that let Microsoft store submitted data for abuse monitoring. For calls carrying financial or health details, that is a compliance problem, not a footnote.

Microsoft is also not the cheapest transcriber. Its 2.5% word error rate edges Grok Voice Transcribe 2.0 at 2.73%, but at 2.7 times Grok's \$0.20 per hour. Benchmark error rates are a ceiling. Tech Times cites academic work finding real-world error rates above 13% for systems that test under 5%.

Two of the vendor claims need correcting. ElevenLabs' preference figures come from its own blind tests. And its own launch post describes the

promotion as "3x credits" for Creator plans and above, which works out to roughly a two-thirds cut, while Digital Applied's per-character API rates imply 72%. We report both rather than reconcile them.

FIG. 2: ESTIMATED COST PER CONVERSATION-HOUR, SPEECH LAYER AGAINST HUMAN WAGE



BARS SCALED TO THE WAGE (\$21.53). ESTIMATES ASSUME THE AGENT SPEAKS FOR 30 MINUTES OF EACH HOUR AT ABOUT 1,000 CHARACTERS PER MINUTE (DIGITAL APPLIED'S RATIO), 25 AUDIO TOKENS PER SECOND FOR GEMINI, AND STREAMING TRANSCRIPTION FOR THE FULL HOUR AT \$0.54. WAGE EXCLUDES BENEFITS; SPEECH COST EXCLUDES THE LANGUAGE MODEL, TELEPHONY AND PLATFORM FEES. SOURCES: BLS; DIGITAL APPLIED; MICROSOFT AI.

IV. The economic ripple runs through prices, not yet payrolls

Prices. On the assumptions in Fig. 2, the voice layer of an agent costs roughly \$1 to \$3 per conversation-hour at list prices, against the \$21.53 median hourly wage for US customer service representatives. That is an estimate, and it leaves out the larger costs: language-model tokens, telephony, integration and supervision. But it means that halving or doubling TTS prices barely moves the business case. The introductory discounts matter for developer adoption, not for the labour arithmetic.

Labour. The exposed pool is large. BLS counts 2.67 million US customer service jobs and projects a 5% decline over 2025–35, citing automation. Gartner put the global agent workforce at about 17 million when it predicted in 2022 that conversational AI would cut agent labour costs by \$80 billion in 2026. The fine print in that forecast was that only one interaction in ten would be fully automated.

Survey evidence since then cuts against fast displacement. In a survey of 321 service leaders in October 2025, only 20% reported AI-driven headcount reductions. 55% held staffing flat while handling more volume. Gartner separately predicts that half of organisations planning major service-workforce cuts will abandon those plans by 2027.

Market structure. Cheap speech hurts vendors who sell only speech. Transcription is the most exposed: Microsoft's batch model costs \$0.10 per hour, and rivals stream for \$0.18 to \$0.20. Value shifts to whoever owns the agent platform, the customer relationship and the telephony. ElevenLabs is already moving there with ElevenAgents. Single-model API vendors without a platform are the likeliest consolidation targets.

Capital. Investors are still betting that specialists win. ElevenLabs raised \$500 million at an \$11 billion valuation, and TechCrunch reports talks at \$22 billion. If hyperscaler bundling were compressing specialist margins, that valuation would be hard to defend. The next funding disclosure is a direct test.

When the voice costs a dollar or two an hour and the human costs twenty, the price of speech has stopped being the reason a call centre still employs people.

ANALYTICAL ESTIMATE FROM OCTOBER 2026 LIST PRICES AND BLS 2025 WAGE DATA. GARTNER'S SURVEY FOUND ONLY 20% OF SERVICE LEADERS HAD CUT HEADCOUNT THROUGH AI.

V. Second-order effects: the bottleneck moves

Once speech is cheap, the constraints that remain are reliability, trust and law. Google's voice cloning is unavailable in Illinois, Texas, the EEA, the UK, Switzerland and India, and all output carries a SynthID watermark. Biometric and consent rules, not model quality, now decide where cloned voices can ship. For a Spanish or European deployer, the cloning feature does not exist yet.

The second effect is that cost moves upstream. With the speech layer near a dollar an hour, the language model handling each turn becomes the dominant variable cost in most voice agents. Labs that own both the model and the voice can price the bundle as one product. This is a structural advantage for

Google and Microsoft over stand-alone voice vendors that pay someone else for reasoning.

The third effect is on offshore service hubs. Countries whose outsourcing industries rest on voice work face the same arithmetic as US call centres, at lower wages. We found no recent primary data sizing that exposure, so we flag it as a direction rather than a figure.

VI. Precedent: Klarna priced the labour, then priced the quality

In February 2024, Klarna said its AI assistant was doing the work of two-thirds of its customer service chats. By May 2025 it was recruiting human agents again, and its chief executive acknowledged that cost had driven the rollout and quality had suffered. The lesson for voice is direct. The cost of the machine was never the binding constraint. The cost of a bad interaction was.

VII. How it plays out: three scenarios

Probabilities are our analytical judgment, not model forecasts, and they sum to 100%.

1 · Baseline: promotions expire, list prices hold ~55%

3-12 MONTHS ·
STEADY

Introductory rates lapse on schedule and list prices settle between about \$15 and \$40 per million characters. The speech layer stays at \$1 to \$3 per conversation-hour, voice agents spread through narrow, scripted call types, and specialists defend share by selling agent platforms. Watch for the ElevenLabs discount ending on 12 October without extension, Microsoft publishing a general-availability price at or above \$0.54, and Google holding the 1 January doubling.

2 · Acceleration: the discounts become the price ~25%

3-9 MONTHS · FAST

A hyperscaler makes its introductory rate permanent or another lab undercuts it, pushing the speech layer below \$0.50 per conversation-hour. Platform bundling squeezes single-model vendors into consolidation, and enterprises move more call types to full automation. Watch for Google cancelling the 1 January increase, ElevenLabs making discounted Turbo pricing permanent, and a rival lab cutting TTS prices before year-end.

3 · Stall: cheap voice, slow deployment ~20% 6-18 MONTHS · SLOW

Prices fall but adoption does not follow, because preview models without SLAs, consent rules and customer pushback cap deployments. Klarna-style reversals multiply and headcount stays flat. Watch for MAI-Transcribe-2-Streaming still in preview in 2027, Gartner's next survey showing no rise from 20% in AI-driven headcount cuts, and new cloning restrictions in EU or US states.

WHAT WOULD FALSIFY THIS THESIS

The thesis is that the price cuts are temporary and that speech cost no longer constrains voice automation. It fails on the first count if two of the three vendors make their introductory rates permanent before February 2027: then this is a genuine price war, not a promotional cycle. It fails on the second count if survey or employment data show a clear jump in AI-driven headcount cuts in customer service during 2027 that coincides with the price cuts. That would mean speech cost was binding after all.

WHAT TO WATCH, IN ORDER

1. **12 October 2026** — ElevenLabs' 72% launch discount ends. An extension would signal real price pressure from Google and Microsoft.
2. **Late October 2026 (expected)** — Microsoft and Alphabet quarterly earnings calls, for any disclosure on voice-agent usage or Foundry and Gemini API speech revenue.
3. **31 December 2026** — Microsoft's introductory pricing for MAI-Transcribe-2-Streaming ends. The general-availability price, and whether an SLA arrives with it, is the key data point.
4. **1 January 2027** — Gemini 3.8 Flash TTS audio output is scheduled to double to \$18 per million tokens. Whether Google holds the increase tests the promotional reading.
5. **Next Gartner customer service survey, expected late 2026 or 2027** — whether the 20% share of leaders reporting AI-driven headcount cuts rises.

SOURCES

1. Google, "Gemini 3.8 Flash TTS and Gemini 3.8 Flash-Lite TTS," 23 September 2026 · blog.google
2. DataNorth, "Google releases Gemini 3.8 Flash TTS," September 2026 · datanorth.ai
3. Digital Applied, "Gemini 3.8 Flash TTS: Voice Cloning and a Price That Doubles," 23 September 2026 · digitalapplied.com
4. ElevenLabs, "Eleven v4: Our most expressive text-to-speech AI model yet," 28 September 2026 (updated 1 October) · elevenlabs.io
5. Digital Applied, "Eleven v4 Turbo for voice agents: latency and pricing," 28 September 2026 · digitalapplied.com
6. TechCrunch, "ElevenLabs' new v4 speech model supports more expression control and 90 languages," 28 September 2026 · techcrunch.com
7. Microsoft AI, "Our first streaming transcription model debuts at no. 1 on Artificial Analysis," 1 October 2026 · microsoft.ai
8. SiliconANGLE, "Microsoft targets ultra-realistic voice agents with its first streaming transcription model," 1 October 2026 · siliconangle.com
9. MarkTechPost, "Microsoft AI Releases MAI-Transcribe-2-Streaming," 2 October 2026 · marktechpost.com
10. Tech Times, "Microsoft Ships First Streaming Transcription Model, Completing Voice Agent Pipeline on Foundry," 2 October 2026 · techtimes.com

11. Rajesh Beri, "Microsoft's MAI-Transcribe-2-Streaming Tops WER With No SLA," 3 October 2026 · beri.net
12. US Bureau of Labor Statistics, "Customer Service Representatives," Occupational Outlook Handbook, 2025 data · bls.gov
13. Gartner, "Gartner Predicts Conversational AI Will Reduce Contact Center Agent Labor Costs by \$80 Billion in 2026," 31 August 2022 · gartner.com
14. Gartner, "Gartner Predicts 50% of Organizations Will Abandon Plans to Reduce Customer Service Workforce Due to AI," 10 June 2025 · gartner.com
15. Gartner, "Gartner Survey Finds Only 20% of Customer Service Leaders Report AI-Driven Headcount Reduction," 2 December 2025 · gartner.com
16. Customer Experience Dive, "Klarna changes its AI tune and again recruits humans for customer service," 9 May 2025 · customerexperiencedive.com

METHOD NOTE: PRICES ARE ATTRIBUTED TO THE VENDOR POST OR, WHERE THE VENDOR DID NOT PUBLISH A RATE, TO THE SECONDARY SOURCE THAT REPORTED IT. GOOGLE'S 1 JANUARY 2027 PRICE INCREASE AND ELEVENLABS' PER-CHARACTER API RATES COME FROM SECONDARY SOURCES READING THE VENDORS' PRICING PAGES; ELEVENLABS' OWN POST DESCRIBES THE PROMOTION DIFFERENTLY, AND BOTH VERSIONS ARE REPORTED. COST-PER-CONVERSATION-HOUR FIGURES ARE OUR ESTIMATES, WITH ASSUMPTIONS STATED UNDER FIG. 2. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.

Built by the [Alicante Tech Vibes](#) community, a tope.

2026-10-04