New Voice Features and UX Developments

Deep Research report · 150 sources

The Pattern: Voice Input as Table Stakes for Text Tools

A wave of productivity and chat tools added basic voice dictation this year, treating it as parity work rather than differentiation:

- **Lattice** added a voice icon to its Agent text field with automatic speech-to-text transcription for hands-free message composition [1].
- **Notion** extended voice input from mobile-only to desktop, letting users capture ideas without typing [44]. It had already added consent controls letting users opt in or out of AI processing for its separate AI Meeting Notes feature, addressing privacy concerns around audio handling [85].
- **xAI's Grok** picked up voice dictation on iOS, routed through three pathways: the Shortcuts app, the Action Button, or Siri invocation, shipped in app version 1.3.97 [46].
- **OpenAI** pushed voice mode to the ChatGPT desktop app (previously web and mobile only), letting users control ChatGPT Work and Codex agents by voice, and separately rolled the same capability out globally on macOS and Windows for Plus through Enterprise tiers [91][105].

The underlying pattern across all four: speech-to-text is now assumed infrastructure, and the interesting differentiation has moved to what voice lets you control (agents, multi-step tasks) rather than the transcription itself.

The Technology Underneath: New Voice Models and Latency Gains

- **OpenAI's GPT-Live** represents the most significant model-level voice upgrade of the period. It now powers ChatGPT Voice and is built for more fluid, human-like conversational exchange, with the architecture emphasizing reduced latency and improved prosody rather than just transcription accuracy [16]. This model is also the engine behind OpenAI's voice-driven agent control features on desktop [91][105].
- **HeyGen** cut LiveAvatar session startup time from roughly 2.2 seconds to 1.2 seconds, driven by a signaling-latency drop from about 950ms to about 12ms, a roughly 45% improvement in initialization speed that matters directly for real-time customer-facing use cases like live event coverage [20].
- **NVIDIA's Magpie TTS**, distributed as open weights via Hugging Face, expanded to 12 languages (adding Modern Standard Arabic, Korean, and Brazilian Portuguese) with Time to First Audio as low as 32ms on an NVIDIA B200 GPU and 79ms on A100 hardware, deployed as an on-premises NIM container to avoid managed-service round-trip latency [3]. This is notable because it targets

teams that want data residency and latency predictability over API simplicity, an architectural choice distinct from cloud-hosted voice APIs.

- **Cohere** released Transcribe, an open-source speech recognition model positioned as matching or beating proprietary speech-to-text systems, aimed at developers who want to self-host rather than depend on a cloud speech API[33].
- **Google DeepMind's Gemini Omni** processes text, audio, and video as native inputs and outputs without converting between modalities, which the company frames as reducing latency and improving reasoning coherence for real-time audio and video tasks[129]. Google separately shipped
- **Gemma 4 12B**, a multimodal model that merges text and image processing into a single architecture without a separate image encoder, cutting computational overhead[146].

Conversational and Agentic Voice UX in Production

- **Twilio's Conversation Relay** is the clearest example of voice UX innovation aimed at eliminating the "robotic" feel of voice AI: it now handles mid-sentence interruptions (letting callers cut off the AI) and inserts natural pauses to reduce mechanical cadence, demonstrated in a tutorial combining Twilio, OpenAI, and PHP[140]. Twilio's broader platform bundles this with Conversation Memory for persistent context and Conversation Orchestrator for coordinating human/AI handoffs, built on a tech stack that mixes OpenAI, Anthropic, Deepgram, ElevenLabs, Google Cloud Speech-to-Text, and Amazon Polly depending on customer choice [121][134] .
- **Palantir** added voice-integrated agent interaction to its Ontology Foundations platform at DevCon 5, paired with what-if scenario modeling and granular security controls, explicitly framed around making agents production-ready rather than experimental[12].
- **OpenAI Presence**, a new enterprise platform for deploying voice and chat agents in both customer-facing and internal workflows, signals OpenAI moving from API access into packaged agent products that compete directly with dedicated enterprise automation vendors[31].
- **Intercom** rebuilt its voice assistant as Fin Voice 2, adding more than 20 features and, notably, moving it onto Apex Flash, Intercom's own proprietary foundation model built for low-latency voice rather than a third-party LLM [61][42] . This is one of the few examples in this set of a company deliberately swapping out third-party model dependency for owned infrastructure specifically for voice.
- **Zendesk** similarly runs voice AI agents integrated with Amazon Connect, using voice to handle routine authentication and resolution tasks so human agents can focus on more complex interactions[27].

Consumer Product Voice Features

- **Spotify** rolled out combined voice and text search to Premium users, letting them ask questions or make requests directly within the Home and Now Playing views to shape recommendations

and understand what's currently playing [17]. The Premium-only gating suggests Spotify is testing willingness to pay for conversational personalization rather than treating it as a free-tier hook.

- Amazon's Alexa for Shopping added conversational search that answers natural-language questions ("where are my AA batteries?") with personalized, context-aware results, and can analyze photos of a customer's fridge or bookshelf to suggest relevant products, described internally as an "agentic AI shopping assistant" [26]. This sits alongside Amazon's broader Alexa+ push as a next-generation voice assistant with agentic, multi-turn reasoning bundled free for Prime members [11]. Separately, Amazon has been engineering Alexa+ to depend less on Anthropic's Claude and optimizing internal GPU utilization, a deliberate move to control AI economics and reduce reliance on third-party frontier models for a flagship voice product [41].
- Suno pushed voice into music creation across several surfaces: an iMessage keyboard extension that lets users create songs by voice or text and send them to contacts who can reply with their own compositions [19], direct import of iOS Voice Memos audio into its Create form [49], and a vocal gender selector plus prompt memory added to its mobile apps to match existing web functionality [79]. Its new Studio 2.0 browser-based DAW adds MIDI recording via a musical typing mode with an arpeggiator, letting users play notes on a keyboard as an alternative production input [70][99] .

Accessibility-Driven Voice and Signing Innovation

- Google DeepMind's SL2T is arguably the most technically ambitious voice-adjacent UX shift in this set: a sign-language-to-text model trained on more than 100,000 hours of data across 50-plus sign languages, launched in Gboard and Live Transcribe on Pixel 11 starting with American Sign Language to English [48] [93]. It uses an on-device MediaPipe Holistic model to track pose landmarks and translates directly from those landmarks to text, discarding raw video immediately and sending only geometric coordinates to the server, a privacy-first architecture that bypasses the intermediate "gloss" annotation step other sign-language systems have relied on [48]. Google built it in collaboration with the Deaf community [93].
- Meta twice paused monetization plans for accessibility-adjacent audio features on its Ray-Ban and Meta smart glasses after user backlash: a planned $20/month subscription for an on-device audio clarity feature [55], and a rate-limiting plan for Conversation Focus, an on-device accessibility feature [59]. Both features run locally without cloud dependency, and Meta says it remains interested in subscription models for other features even as it backed off these two [59].

Enterprise Voice Agent Platforms at Scale

- ElevenLabs has moved well beyond text-to-speech into a full agent and creative platform. Its
- ElevenAgents product now handles more than 10 million conversations weekly across tasks like refunds, bookings, and benefits inquiries, with simulation, A/B testing via a "Spotlight" monitoring tool, and enterprise features like configurable guardrails and regional data residency [76]. It further

expanded agent deployment channels to SMS, Telegram, Intercom, and Freshdesk, alongside existing phone, web, Zendesk, Slack, and WhatsApp support, and added Singapore data residency and a Canada launch for compliance-sensitive customers [128]. On the voice-cloning side, ElevenLabs added a Professional Voice Clone feature that generates AI voices from uploaded audio or fresh microphone recordings in a three-step workflow [118] [145], and a separate voice-swap feature that applies any of more than 10,000 library voices to a user's own recorded delivery while preserving original pacing and intonation [150].

- ElevenLabs' pricing treats voice generation as a metered resource from the start: even its free tier bundles Text to Speech, Speech to Text, Sound Effects, and Voice Design at $0, with usage billed per minute and rates dropping from about $0.36/minute on Free to about $0.17/minute on the Business tier [6].

Infrastructure Behind Real-Time Voice and Video Calling

Several companies invested in the plumbing that makes voice and video interaction feel instantaneous:

- Discord migrated its voice and video infrastructure from centralized data centers to Cloudflare's distributed edge network specifically to cut latency by routing traffic through geographically closer servers [35].
- Meta deployed the AV1 video codec across its real-time communication platforms after a multi-year effort addressing codec selection, device eligibility, rate control for variable network conditions, and error resilience, developing proprietary rate-control technology along the way [50].
- Meta's WhatsApp Web gained audio and video calling with no app download required, call transfer between devices without disconnecting, QuickHD video that reaches high definition within seconds, and toggleable noise suppression [109].
- SoundCloud replaced its audio encoder for the first time in more than a decade, switching to Fraunhofer's libfdk_aac to improve quality at the same bitrate while shrinking file sizes, and retroactively re-encoded its existing catalog [25][40] .

Avatar and Video-as-Voice-Interface

- HeyGen's LiveAvatar and related products extend "voice" into full conversational video presence. Its April 2026 release included Avatar V, generating full upper-body movement and multiple looks from a single 15-second recording, and a CLI tool that lets AI agents generate video directly through structured JSON commands, auto-selecting avatar, voice, script, and layout [9]. Its May release added Custom Motion, letting users direct avatar performance (gaze, gestures, posture, energy) through plain-English instructions, priced at $0.05 per second via API [24]. HeyGen co-founder Wayne Liang used the company's own avatar technology to handle sales calls during his parental leave, a real internal test of whether avatar substitution can hold up in a business-critical, customer-facing role [54].

- **Synthesia** launched Roleplay Sessions, interactive avatars that simulate realistic pushback for practicing conversations like salary negotiations, with built-in scoring and feedback, moving the company from passive video generation into live, two-way interactive training [39] [69] [112]. It also shipped Dubbing 2.0, combining automatic lip-sync with voice cloning to translate video into other languages, offering 15 minutes of free daily dubbing through a limited promotion [94].

The Open Question: Is Voice a Feature or the Interface

- Apple CEO Tim Cook signaled that the company is considering a tiered compute model for Siri, where enhanced AI capabilities would be paid upgrades delivered through existing iCloud+ subscriptions, though no pricing or timeline was given [34]. Apple is separately negotiating multiyear publisher licensing deals, potentially worth nine figures, to feed Siri real-time news rather than relying on static training data, directly responding to the fact that competing LLM assistants already have web access [108]. Apple's tech stack includes Siri Expressive Voices, on-device speech synthesis powered by its AFM 3 Core Advanced model and a memory-efficient diffusion architecture [64].
- Toast has posted openings for both a Senior and a Staff Software Engineer on its Voice AI Platform team [47] [62], and xAI has a listing for an AI Tutor focused on Audio Editing [66], both suggesting continued internal investment in voice capability building even where public product announcements haven't yet followed.

Sources

Launch

9 Launched Avatar V, Seedance 2.0 integration, Instant Highlights v2, and Video Agent CLI in April

12 Added voice control and scenario modeling to make AI agents production-ready

16 Released GPT-Live voice models enabling lower-latency ChatGPT speech

24 Shipped 18 features in May: HyperFrames, Avatar V upgrades, LiveAvatar, Android app, four

29 Added summarization and translation APIs to developer platform

31 Launched OpenAI Presence, enterprise AI agent platform for voice and chat

33 Released open-source speech recognition model competing with proprietary alternatives

48 Launches sign-language-to-text model in Gboard and Live Transcribe, starting with ASL

61 Switched voice AI to proprietary foundation model with 20+ new features

63 Released DiffusionGemma, a diffusion-based text model 4x faster than autoregressive generation

67 Released Command A Vision, a multimodal model accepting text and image inputs

69 Launched AI Roleplay Sessions for interactive employee training beyond video

76 Launched ElevenAgents platform for building and deploying conversational AI at scale

81 Released 1000+ translation models from University of Helsinki

93 Launches SL2T sign-language-to-text model on Pixel 11 and Android

96 Released Muse Glimmer, 30B multimodal model for local agentic tasks

99 Launched Studio 2.0 browser-based DAW with granular production controls

115 Launched Gong Revenue Harness, orchestration layer for multi-agent revenue workflows

124 Launches HeyGen for Real Estate, three video formats for agents

128 Launched agents for SMS, Telegram, Intercom; added character voices to audiobooks and song

129 Gemini Omni processes text, audio, video natively without conversion latency

132 Lambda one-click setup prompt, OpenAI GPT-5.6 models added to Bedrock

146 Merged text and image processing into single 12B model, eliminating separate encoder

Feature

1 Agent voice input added to Lattice with automatic transcription

2 Redesigned workspace menu with improved navigation and mobile support

3 Magpie TTS expands to 12 languages with Arabic, Korean, Portuguese support

5 Call Logs APIs now return handler details across phone endpoints

7 Integrated Google's Gemini Omni Flash into ElevenCreative for unified video creation and editing

10 Brave 1.93 masks GPU fingerprinting via WebGL de-identification and extension noise

17 Rolled out voice and text search to Premium users for personalized recommendations

19 Added iMessage keyboard integration for song creation and sharing

20 LiveAvatar session startup latency cut from 2.2s to 1.2s

21 Flows agent automates documentary video creation from script to final cut

22 Integrated Gemini Omni Flash into video editor for object editing and image animation

26 Alexa for Shopping adds conversational search with personalized recommendations and image

34 Cook signals paid tier for advanced Siri AI features via iCloud+

37 Integrated Google's Gemini Omni video generation into platform

39 Launched Roleplay Sessions tool for interactive avatar practice scenarios

44 Adds voice input to desktop version of Notion

46 Grok gains voice dictation via iOS shortcuts and Siri

49 iOS app gains direct import from Notes and Voice Memos

51 Seedream 5.0 Pro adds layer-based editing and native 14-language text generation

52 Integrated Gemini video model enables eight text-prompt edits while preserving character

55 Pauses paid subscription plan for Ray-Ban smart glasses audio feature

59 Pauses rate-limiting plan for Conversation Focus accessibility feature on Meta glasses

65 Added SMS deliverability check to prevent wasted messages to invalid numbers

70 Studio 2.0 adds MIDI editing, audio effects, synths, and automation to web DAW

77 Mobile app redesigned navigation bar around user identity and preview upcoming features

79 Mobile app gains vocal gender selection, lyrics model picker, prompt memory

80 Launched AI Agent to eliminate response-time delays in sales outreach

82 Added audio, video transcription and enhanced image understanding to core product

84 Launched AI tool that auto-generates motion graphics directly from scripts

85 Added consent controls for AI Meeting Notes feature

86 Deployed language models to improve product search and recommendations

91 Voice mode now available in ChatGPT desktop app with agent control

94 Dubbing 2.0 adds realistic lip-sync and voice cloning for video translation

100 Introduced composable AI layer for Slackbot to integrate third-party models

101 Retail store wait times fell 13.9% after pairing AI agents with human staff

105 Rolled out voice control for ChatGPT Work and Codex across desktop

109 WhatsApp Web gains calling, call transfer, and noise suppression

110 Integrated Google's Gemini Omni model for generating AI videos with personal avatars

113 Switched streaming APIs to AAC HLS codec for improved audio quality

118 Professional Voice Clone feature lets users create AI voices from existing content or mic

125 Published apps can now integrate with AI assistants like ChatGPT and Claude

130 Call recordings now available on mobile apps

134 Conversation Relay enables real-time AI phone agents via Twilio

138 AI Gateway made available through AWS Marketplace

140 Conversation Relay now handles interruptions and natural pauses in AI voice

141 Added encoder-decoder architecture support to Transformers library via hybrid combiners

145 Added voice cloning to ElevenCreative creative suite

150 Added voice-swap feature letting users apply library voices to their own delivery

Internal tool

25 Upgraded AAC audio encoder for first time in over a decade

35 Migrated voice and video infrastructure to Cloudflare edge network for lower latency

36 How I taught Claude to write Maestro tests (so I don't have to)

40 Upgraded AAC encoder for first time in over a decade to improve audio quality

50 Rolled out AV1 video codec to real-time calls after multi-year infrastructure rebuild

56 Buzz uses MeshLLM to pool idle compute for distributed LLM inference

71 Built multi-cloud infrastructure over three years to run enterprise AI models

72 Built internal Slack bot that analyzes customer calls nightly to surface themes and quotes

74 Code Search: How Agents search across Snap's codebase

103 Rebuilt accessibility features for canvas-based design tool

116 Tactical Coding Assistants

120 Built machine-learning system to route LLM requests to fastest model per request

136 Debugged multi-stage latency cascade in usage estimation service across network proxy, kernel,

AI adoption

54 Co-founder deployed AI avatar to handle sales calls during parental leave

87 Replaced rigid menu-tree chatbot with conversational AI for support

108 Negotiates multiyear publisher deals worth nine figures to feed current news into Siri AI

148 Bezos directs Prime Video overhaul centered on AI via internal Project Lighthouse effort

Org

47 Hiring Senior Software Engineer, Voice AI Platform

62 Hiring Staff Software Engineer, Voice AI Platform

Research

Perspective