Newsletter (19) - Voice AI

(1) Introduction

In this week's edition of our newsletter, we explore the technology behind Voice AI, systems that generate, recognize, and interact through spoken language, and explain how companies, particularly ElevenLabs, are building and scaling in this emerging field. Over the past decades, human-computer interaction has followed a clear trajectory: from keyboards to graphical user interfaces, to text-based digital assistants. With the rise of LLMs, text became the primary interface for AI systems. Voice, however, marks the next logical step, enabling a more natural, expressive, and scalable form of interaction.

(2) Technology behind Voice AI

The increased use of Voice AI is associated with development in model architectures, higher availability of computing resources, and broader enterprise development. By automating parts of audio generation, enterprises can reduce manual production effort and standardize voice output across media, entertainment, education, and related content-driven applications.

From a technical perspective, Voice AI systems are fundamentally composed of three integrated components working in sequence: automatic speech recognition (ASR) models that convert audio signals into text, language models that process linguistic structure and context, and speech synthesis models that generate audio output. These components form a processing pipeline where quality is determined less at the final synthesis stage and more during initial audio analysis and interpretation. Speech recognition, therefore, serves as a foundational layer across nearly all Voice AI applications, including dubbing, which enables natural multilingual speech, voice cloning, long-form narration, and conversational agents. A language model then interprets this information, inferring linguistic structure, intent, and context, after which timing and prosody are adapted to align speech with narrative constraints. Finally, a vocoder converts the processed representation into an audible waveform. In practice, this process begins with converting audio into spectrograms, which transformer-based neural networks use to map sound patterns to language.

In practice, these architectural principles are reflected in how ElevenLabs designs its voice AI systems. ElevenLabs treats speech recognition as part of an integrated system where listening, understanding, timing, and synthesis are optimized jointly. High-quality, expressive training data enables models to capture intonation, pacing, and emotional nuance. Proprietary speaker diarization, identifying who speaks when, allows individual audio segments to be edited without disrupting timing or tonal consistency. Together, these system-level capabilities form an audio infrastructure that is difficult to replicate at scale and becomes increasingly valuable in multilingual and long-form use cases.

(3) Founding Story of Elevenlabs

Understanding the technical architecture underlying Voice AI provides context for how ElevenLabs emerged as a leader in this space, driven by founders who recognized early limitations in existing speech synthesis technology.

ElevenLabs was founded in April 2022 by Piotr Dąbkowski and Mati Staniszewski, who first met as teenagers at Copernicus High School in Warsaw. Staniszewski studied mathematics at Imperial College London before working at Palantir, while Dąbkowski pursued AI and machine

learning at Oxford and Cambridge, publishing research at NeurIPS. Before founding ElevenLabs, they collaborated on several projects, including an accent-detection application and recommendation engine. Throughout these projects, they were consistently frustrated by the robotic quality of existing text-to-speech technologies, including widely used systems like Siri and Alexa.

Their motivation to solve this problem was further reinforced by their shared experience with Poland's film dubbing industry, where a single actor traditionally voiced every character in foreign films, resulting in monotone audio lacking emotional variation. This research led them to develop methods for context awareness and emotional intelligence in speech synthesis. Early prototypes successfully replicated natural pauses, laughter, and conversational fillers, demonstrating that lifelike artificial voices were technically feasible.

The pre-seed fundraising proved to be challenging, as investors questioned market size and defensibility against tech giants. After pitching to approx. 50 investors, they secured backing from venture capital funds Credo Ventures and Concept Ventures, as well as from angel investors including Polkadot co-founder Peter Czaban. They strategically delayed announcing funding until the January 2023 beta launch of their text-to-speech platform, which rapidly attracted over one million users within five months.

(4) Success Story

Following strong adoption among creators, ElevenLabs quickly translated early usage into paid demand across both self-serve and enterprise customers. Its success has been driven less by hype cycles than by consistent, rapid execution. After the launch of its public beta, the company scaled from zero to $100 million in annual recurring revenue (ARR) in 20 months, followed by $200 million ARR in the next 10 months, and reached $330 million ARR in just five additional months by early 2026, placing ElevenLabs among the fastest-scaling AI companies globally. This commercial traction was mirrored in its fundraising trajectory. Having raised a difficult $2m pre-seed at a $9m post-money valuation in 2022, ElevenLabs' subsequent $180M Series C in January 2025 at a $3.3 billion valuation was completed at roughly 30x ARR, signaling strong investor conviction in both durability and upside.

By late 2024, ElevenLabs had transitioned from a creator-first product into an enterprise-grade platform built around long-term contracts, high average contract values, and increasingly predictable cash flows. In January 2026, Deutsche Telekom partnered with ElevenLabs to deploy AI voice agents within customer service operations across app and phone channels. Similarly, leveraging Elevenlabs' technology, Revolut rolled out voice agents across the UK and Europe, with the system reducing time to resolution by 8x while serving 4 million customers in more than 31 languages and achieving call success rates of 99.7%. Beyond customer-facing use cases, TELUS Digital integrated ElevenLabs' technology into its agent training platform, where AI voice simulations enhanced onboarding efficiency and reduced employee turnover, demonstrating early evidence of structurally lower churn.

(5) Drivers of Success

ElevenLabs's success stems primarily from maintaining clear product superiority through emotion-rich speech synthesis and multilingual support across 32 languages. Their research models consistently outperform OpenAI and Google on multiple benchmarks. According to ElevenLabs, their speech-to-text model achieved higher accuracy rates across 99 languages than Google's Gemini 2.0 Flash and OpenAI's Whisper Large V3; this advantage exists largely because ElevenLabs was among the first to focus exclusively on Voice AI, building proprietary

methods for context awareness and emotional intelligence that larger labs, attempting to solve multiple modalities (text, image, video, voice simultaneously), have struggled to match. The company rapidly diversified from text-to-speech into a comprehensive suite, including Projects for long-form audio, Dubbing Studio for video localization, Audio Native for website integration, the speech-to-text model Scribe, and ElevenReader for mobile applications.

The explosion of agentic AI drives primary demand, with the majority of revenue coming from their conversational AI platform. Their clients span 41% of Fortune 500 companies, including Cisco, Nvidia, and Adobe. The audiobook market provides another growth vector, projected to grow from $8.7bin 2025 to $35.5bby 2030. The ElevenReader Publishing suite democratizes audiobook production, while partnerships with HarperCollins, one of the world's largest trade publishers, video editing platform Kapwing, and media conglomerate Bertelsmann amplify market reach.

The October 2024 acquisition of Omnivore, a startup specializing in automated voice pipelines for media distribution, strengthened media localization capabilities. Unlike competitors relying on third-party models, ElevenLabs controls research, product, and distribution through vertical integration, providing superior unit economics. This allows their system to respond within under 100 milliseconds, which is faster than the majority of competitors. Furthermore, they also focus on making voices sound more human by training on large speech datasets they collected themselves, helping the AI produce natural pauses, laughter, and a realistic speaking rhythm instead of relying only on general AI models.

(6) Conclusion

In conclusion, the Voice AI market is experiencing a structural transformation as model performance improves, enterprise use cases expand, and voice emerges as a primary mode of human-computer interaction. As deployment moves from experimentation to adoption, companies are integrating voice systems into customer service, media production, and operational workflows. ElevenLabs has positioned itself at the center of this transformation through focused research in voice synthesis, rapid product diversification, and early enterprise traction. Companies that control research, product, and distribution will be best positioned to capture sustained value in this emerging infrastructure layer. From a venture capital perspective, investment in Voice AI surged from approximately $315 million in 2022 to $2.1 billion in 2024, a seven-fold increase in two years, signaling that institutional investors view the category as entering a platform cycle comparable to early cloud infrastructure or mobile interfaces.