Now You’re Speaking My Language: NVIDIA Riva Sets New Bar for Fully Customizable Speech AI

Author: Scott Martin

Whether for virtual assistants, transcriptions or contact centers, voice AI services are turning words and conversations into bits and bytes of business magic.

At GTC this week, NVIDIA announced new additions to NVIDIA Riva , a GPU-accelerated software development kit for building and deploying speech AI applications.

Riva’s pretrained models are now offered in seven languages, including French and Hindi. Additional languages on the horizon: Arabic, Italian, Japanese, Korean and Portuguese. Riva also brings improvements in accuracy for English, German, Mandarin, Russian and Spanish. Additionally, it adds capabilities like word-level confidence scores and speaker diarization — the process of identifying speakers in audio streams.

Riva is built to be fully customizable at every stage of the speech AI pipeline to help solve unique problems efficiently. Developers can also deploy it where they want their data to be: on premises, for hybrid multiclouds, at the edge or in embedded devices. It’s used by enterprises to bolster services, efficiency and competitive advantage.

While AI for voice services has been in high demand, development tools have lagged. More people are working and learning from home, shopping online and seeking remote customer support, which strains call centers and pushes voice applications to their limits. Customer service wait times have recently tripled as staffing shortages have hit call centers hard, according to a 2022 Bloomberg report.

Advances in speech AI offer the way forward. NVIDIA Riva enables companies to explore larger deep learning models and develop more nuanced voice systems. Speech AI applications built on Riva provide an accelerated path to better services, promising improved customer experiences and engagement.

The worldwide market for contact center software reached about $27 billion in 2021, a figure expected to nearly triple to $79 billion by 2029, according to Fortune Business Insights.

This increase is due to the benefits that customized voice applications offer businesses of any size, in almost every industry — from global enterprises, to original equipment manufacturers delivering speech AI-based systems and cloud services, to systems integrators and independent software vendors.

NVIDIA Riva includes pretrained language models that can be used as is or fine-tuned using transfer learning from the NVIDIA TAO Toolkit , which allows for custom datasets in a no-code environment. Riva automated speech recognition (ASR) and text-to-speech (TTS) models can be optimized, exported and deployed as speech services.

Voice AI is making its way into ever more types of applications, such as customer support virtual assistants and chatbots, video conferencing systems, drive-thru convenience food orders, retail by phone, and media and entertainment. Global organizations have adopted Riva to drive voice AI efforts, including T-Mobile, Deloitte, HPE, Interactions, 1-800-Flowers.com, Quantiphi and Kore.ai.

T-Mobile adopted Riva for its T-Mobile Expert Assist — a custom-built call center application that uses AI to transcribe real-time customer conversations and recommend solutions — for 17,000 customer service agents. T-Mobile plans to deploy Riva worldwide soon.

Hewlett Packard Enterprise offers HPE ProLiant servers that include NVIDIA GPUs and NVIDIA Riva software in a system capable of developing and running challenging speech AI and natural language

processing workloads that can easily turn audio into insights. HPE ProLiant systems and NVIDIA Riva form a world-class, full-stack solution for running financial services and other industry applications.

“To deliver the capabilities of NVIDIA Riva, HPE offers a Kubernetes-based NLP reference architecture based on HPE Ezmeral software,” said Scott Ramsay, vice president of HPE GreenLake solutions at HPE. “Delivered through the HPE GreenLake cloud platform, this system enables developers to accelerate the development and deployment of next-generation speech AI applications.”

Deloitte supports clients looking to deploy ASR and TTS use cases, such as for order-taking systems in some of the world’s largest quick-order restaurants. It’s also developing chatbot services for healthcare providers that will enable accurate and efficient transcriptions for patient questions and chat summarizations.

“Advances in natural language processing make it possible to design cost-efficient experiences that enable purposeful, simple and natural customer conversations,” said Christine Ahn, principal at Deloitte US. “Our clients are looking for a streamlined path to conversational AI deployment, and NVIDIA Riva supports that path.”

Interactions has integrated Riva with its Curo software platform to create seamless, personalized engagements for customers in a broad range of industries that include telecommunications, as well as for companies such as 1-800-Flowers.com , which has deployed a speech AI order-taking system.

Kore.ai is integrating Riva with its SmartAssist speech AI contact-center-as-a-service, which powers its BankAssist, HealthAssist, AgentAssist, HR Assist and IT Assist products. Proof of concepts with NVIDIA Riva are in progress.

Quantiphi is a solution-delivery partner that is developing closed-captioning solutions using Riva for customers in media and entertainment, including Fox News. It’s also developing digital avatars with Riva for telecommunications and other industries.

Speech AI pipelines can be complex and require coordination across multiple services. Microservices are required to run at scale with ASR models, natural language understanding, TTS and domain-specific apps. NVIDIA GPUs are ideal for acceleration of these types of specialized tasks.

Riva offers software libraries for building speech AI applications and includes GPU-optimized services for ASR and TTS that use the latest deep learning models. Developers can meld these multiple speech AI skills within their applications.

Developers can easily access Riva and pretrained models through NVIDIA NGC , a hub for GPU-optimized AI software, models and Jupyter Notebook examples.

Support for Riva is available through NVIDIA AI Enterprise , a cloud-native suite of AI and data analytics software that’s optimized to enable any organization to use AI. It’s certified to deploy anywhere — from the enterprise data center to the public cloud — and includes global enterprise support to keep AI projects on track.

Try NVIDIA Riva with guided labs on ready-to-run infrastructure in NVIDIA LaunchPad.

Original URL: https://blogs.nvidia.com/blog/2022/09/21/riva-speech-ai/