Speech AI is powering conversational AI applications.

Companies across all industries interact with customers for billions of minutes every day. To use insights from these conversations and build better products, companies deploy conversational applications, such as customer care agent assists, virtual assistants, and digital avatars. These applications also need to enable interactions with users in many languages, understand industry-specific jargon, and respond in real time.

Delivering world-class speech AI skills.

NVIDIA Riva is a GPU-accelerated speech AI SDK with automatic speech recognition (ASR) and text-to-speech (TTS) skills for conversational applications in the cloud, on premises, at the edge, and in embedded devices. Riva offers out-of-the-box (OOTB), state-of-the-art speech models that are trained for millions of hours on thousands of hours of audio data. The ASR and TTS pipelines are optimized for real-time performance, with inference running far below the natural conversation threshold of 300 milliseconds. To achieve the best possible accuracy, developers can further fine-tune Riva models on their domain-specific data.

Key Challenges

- **Complexity** – Building speech AI applications requires hundreds of thousands of hours of audio data, tools, and expertise to build and customize models.
- **Accuracy** – Low accuracy leads to a poor customer experience.
- **Performance** – Natural conversation requires responses under 300 milliseconds.

NVIDIA Riva Benefits

- **High OOTB accuracy** – Riva’s pre-trained models enable immediate world-class accuracy.
- **Easy customization** – For domain-specific data, users can fine-tune Riva speech models to achieve the best possible accuracy.
- **Real-time performance** – Riva delivers interactive responses in less than 300 milliseconds and less than 100 milliseconds on embedded devices.
- **High throughput** – Riva’s GPU acceleration provides 7X higher throughput than CPU-based applications.
- **High scalability** – Riva is fully containerized and can easily scale to hundreds and thousands of real-time streams.
- **Flexible deployment** – Riva speech skills run as a service in the cloud, on premises, embedded, and at the edge. They can also be run in real-time streaming and offline batch mode.
Offering high quality automatic speech recognition.

Every speech AI application relies on converting human voice into readable text (i.e., ASR). Often the first step of the speech pipeline, ASR quality influences the effectiveness of downstream conversational AI tasks. Riva offers world-class, OOTB ASR models in several languages—including English, Spanish, German, Russian, Mandarin, Hindi, and French, with Arabic, Korean, Portuguese, and Italian coming soon—empowering enterprises to deploy high-quality speech AI applications globally. These models can be further customized through word boosting, customized punctuation, and inverse text normalization. Models can also be fine-tuned for custom jargon, domain-specific words, and noisy environments.

Creating expressive, human-like text-to-speech.

Generating human voices from text is important for customer-facing services in conversational applications. Producing an expressive and engaging human-like voice requires state-of-the-art AI models, is compute-intensive, and calls for a mature pipeline to fine-tune and express voices. NVIDIA Riva provides professional female and male OOTB voices and enables users to easily customize the models and pipeline for voice pronunciation, pitch, volume, and speed.
Get NVIDIA Riva with enterprise-level support.

For fast and easy development of customer-facing, real-time conversational applications that need to run at scale, NVIDIA Riva offers ready-to-use state-of-the-art models and fully customizable and accelerated pipelines with full enterprise-grade support through NVIDIA AI Enterprise software. In addition to this support, NVIDIA AI Enterprise customers get access to NVIDIA AI experts and instructor-led workshops, priority access to new features, prioritized bug fixes, and maintenance releases.

![Diagram of NVIDIA Riva features]

- Unlimited Usage
  - Scale up to on-premises and any cloud as workloads expand

- Access to NVIDIA AI Experts
  - Get help from 8-5, including access to engineering, and guidance on configuration and performance

- Control Maintenance and Upgrade Schedule
  - Designated software branches can receive long-term support

- Ticket Prioritization
  - Access the latest security fixes and maintenance releases

Figure 4. NVIDIA Riva with NVIDIA AI Enterprise Support: Through NVIDIA AI Enterprise Software with the assurance of enterprise-grade AI experts.

To learn more about NVIDIA Riva, visit: www.nvidia.com/en-us/ai-data-science/products/riva/