Speech AI Expands Global Reach With Telugu Language Breakthrough

Built with the NVIDIA NeMo framework, automatic speech recognition model tops accuracy leaderboards for a competition hosted by IIIT-Hyderabad.

Author: Angie Lee

More than 75 million people speak Telugu, predominantly in India's southern regions, making it one of the most widely spoken languages in the country.

Despite such prevalence, Telugu is considered a low-resource language when it comes to speech AI . This means there aren't enough hours' worth of speech datasets to easily and accurately create AI models for automatic speech recognition (ASR) in Telugu.

And that means billions of people are left out of using ASR to improve transcription, translation and additional speech AI applications in Telugu and other low-resource languages.

To build an ASR model for Telugu, the NVIDIA speech AI team turned to the NVIDIA NeMo framework for developing and training state-of-the-art conversational AI models. The model won first place in a competition conducted in October by IIIT-Hyderabad, one of India's most prestigious institutes for research and higher education.

NVIDIA placed first in accuracy for both tracks of the Telugu ASR Challenge , which was held in collaboration with the Technology Development for Indian Languages program and India's Ministry of Electronics and Information Technology as a part of its National Language Translation Mission.

For the closed track, participants had to use around 2,000 hours of a Telugu-only training dataset provided by the competition organizers. And for the open track, participants could use any datasets and pretrained AI models to build the Telugu ASR model.

NVIDIA NeMo-powered models topped the leaderboards with a word error rate of approximately 13% and 12% for the closed and open tracks, respectively, outperforming by a large margin all models built on popular ASR frameworks like ESPnet, Kaldi, SpeechBrain and others.

"What sets NVIDIA NeMo apart is that we open source all of the models we have — so people can easily fine-tune the models and do transfer learning on them for their use cases," said Nithin Koluguri, a senior research scientist on the conversational AI team at NVIDIA. "NeMo is also one of the only toolkits that supports scaling training to multi-GPU systems and multi-node clusters."

The first step in creating the award-winning model, Koluguri said, was to preprocess the data.

Koluguri and his colleague Megh Makwana, an applied deep learning solution architect manager at NVIDIA, removed invalid letters and punctuation marks from the speech dataset that was provided for the closed track of the competition.

"Our biggest challenge was dealing with the noisy data," Koluguri said. "This is when the audio and the transcript don't match — in this case you cannot guarantee the accuracy of the ground-truth transcript you're training on."

The team cleaned up the audio clips by cutting them to be less than 20 seconds, chopped out clips of less than 1 second and removed sentences with a greater-than-30 character rate, which measures characters spoken per second.

Makwana then used NeMo to train the ASR model for 160 epochs, or full cycles through the dataset, which had 120 million parameters.

For the competition’s open track, the team used models pretrained with 36,000 hours of data on all 40 languages spoken in India. Fine-tuning this model for the Telugu language took around three days using an NVIDIA DGX system , according to Makwana.

Inference test results were then shared with the competition organizers. NVIDIA won with around 2% better word error rates than the second-place participant. This is a huge margin for speech AI, according to Koluguri.

“The impact of ASR model development is very high, especially for low-resource languages,” he added. “If a company comes forward and sets a baseline model, as we did for this competition, people can build on top of it with the NeMo toolkit to make transcription, translation and other ASR applications more accessible for languages where speech AI is not yet prevalent.”

“ASR is gaining a lot of momentum in India majorly because it will allow digital platforms to onboard and engage with billions of citizens through voice-assistance services,” Makwana said.

And the process for building the Telugu model, as outlined above, is a technique that can be replicated for any language.

Of around 7,000 world languages, 90% are considered to be low resource for speech AI — representing 3 billion speakers. This doesn’t include dialects, pidgins and accents.

Open sourcing all of its models on the NeMo toolkit is one way NVIDIA is improving linguistic inclusion in the field of speech AI.

In addition, pretrained models for speech AI, as part of the NVIDIA Riva software development kit, are now available in 10 languages — with many additions planned for the future.

And NVIDIA last month hosted its inaugural Speech AI Summit , featuring speakers from Google, Meta, Mozilla Common Voice and more. Learn more about “Unlocking Speech AI Technology for Global Language Users” by watching the presentation on demand.

Get started building and training state-of-the-art conversational AI models with NVIDIA NeMo .

Original URL: https://blogs.nvidia.com/blog/2022/12/02/speech-ai-telugu-language-breakthrough/