Solution Brief

Speech-Driven Interfaces
Multilingual Speech Recognition

SIL International Automatically Identifies Language and Accent for Touchless Interactions

Edge-ready solution built using Intel® Distribution of OpenVINO™ toolkit and Intel® DevCloud for the Edge supports multilingual transportation and business hubs

"Our new approach uses dynamic loading of monolingual models to achieve speech recognition with high performance across languages and accents. Our hope with this project is to build speech technology that enables developers to extend the benefits of speech-related AI to emerging markets for the post-COVID world."

— Daniel Whitenack, data scientist, SIL International

Up to 75% reduced inference time and memory usage[1]

Up to 90% reduced storage usage[1]

Speech-driven interfaces have experienced a surge of interest as businesses seek solutions that enable touchless interaction for customers during the COVID-19 pandemic. For ticketing and banking kiosks, especially those located in transportation and business hubs, multilingual support is critical to fulfilling customer needs. With its new multilingual speech recognition solution, SIL International makes it possible to recognize languages and accents at the edge.

The solution—prototyped using Intel® DevCloud for the Edge—uses compact neural network architectures and optimizations from the Intel® Distribution of OpenVINO™ toolkit to allow dynamic loading and usage of multiple speech and language models. This reduces the need for compute and memory while creating a completely touchless multilingual experience that does not even require the customer to configure language settings with a touchscreen.

Challenges: Running multiple language models at the edge

Capacity and size limitations create multiple challenges for speech recognition in multiple languages at the edge. Speech and language models are large, and accelerated hardware is often required to run models based on neural networks. When multiple languages are needed, the hardware requirements for running one or more models per language can be prohibitive.

Accents add another layer of complexity to speech recognition. Accent differences as well as demographic factors can dramatically impact the performance of automatic speech recognition (ASR). Typically, these speech variances within a single language are addressed using larger models that use more compute resources or specialized models for particular demographic groups.

This proliferation of models creates a challenge for any kiosk that serves customers using dozens—or even hundreds—of languages and accents. The COVID-19 pandemic has created new urgency around the use of touchless interfaces that use speech recognition technology in international airport, hotel, and banking environments. However, expanding these experiences to new markets and multilingual use cases requires a different approach to language and accent recognition.

See backup for workloads and configurations. Results may vary.

Solution: Dynamic usage and loading for language and accent models

SIL collaborated with Intel to create a system that uses a variety of compact AI models. This system can enable kiosks and other edge devices to support multiple languages and accents. After careful selection of model architectures and optimization using the OpenVINO toolkit, SIL was able to develop integrated AI models that are up to 16x smaller than commonly used pretrained models with minimal degradation in performance.[1]

The solution is also capable of responding to end users in multiple languages, using a natively multilingual natural language understanding (NLU) component.

These new compact integrated models, which do not require kiosks to simultaneously run multiple ASR models, optimize edge device resource consumption. With lower resource consumption and no vendor lock-in, the solution can be run flexibly at the edge, on-premises, or with a cloud provider.

Use cases for the SIL ASR solution include:

- Airport ticketing kiosks
- Transportation system kiosks (buses, trains, etc.)
- Informational kiosks in hotels
- Automated teller machines (ATMs)
- Elevators
- Audio/video analytics
- Videoconference and event captioning
- Retail self-checkout systems

The solution offers built-in, configurable support for multiple languages for both the NLU and ASR components. This enables natural, contactless interactions that automatically identify language and accent for personalized experiences.

How it works

The SIL solution is based on an approach called dynamic ASR, or Dyn-ASR. Audio is first formatted and preprocessed. Next, two models are used: The first, broader model sorts the speaker into a language class. The second model is matched to the language identified in the first model and then classifies speakers by accent.

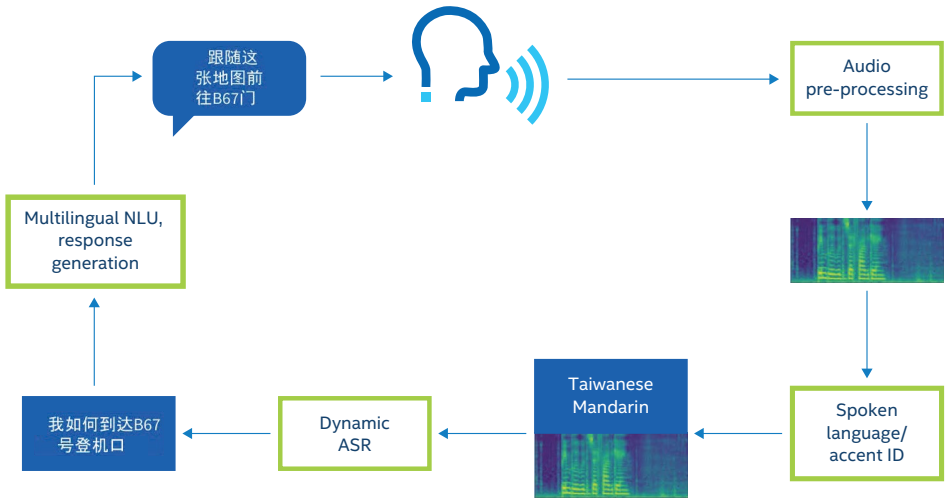
The system uses a separate, tailored ASR model for each supported language and accent pair. ASR models can be loaded into memory and/or used dynamically as the language and accent are identified. This allows multiple models to be used without overtaxing compute or memory resources, allowing the solution to perform at the edge.

SIL used Intel® Core™ i7 and Intel® Core™ i9 processors for its target systems, consistent with the types of processors typically used in the kiosk devices envisioned as its primary use cases. To optimize the ASR models themselves, SIL used a combination of strategic model architecture modifications, fine-tuning with accent-annotated data and the OpenVINO toolkit.

Using Intel DevCloud for the Edge to provide a hosted environment for sandbox solutions, SIL was able to experiment with the toolkit. This allowed SIL to quickly spin up notebooks to test and optimize a variety of models being trained by multiple developers using different environments. Intel DevCloud for the Edge used Jupyter as an interface, a natural fit for SIL's data scientists.

In initial experiments, SIL was able to show that storage and memory usage were reduced by as much as 90 and 75 percent, respectively, for identifying speakers with different accents of US English, Chinese English, mainland Mandarin, and Taiwanese Mandarin, while maintaining a reduction of as much as 75 percent in inference time.[1]

Overview of an example flow of data processing with the SIL solution



See backup for workloads and configurations. Results may vary.

Conclusion: Automatic multilingual, multiaccent speech recognition powered by Intel® technology

Recognizing multiple languages and accents automatically has posed significant challenges for developers of ASR technology, limiting the utility of speech interfaces in international use cases with diverse users. To enable contactless kiosk interfaces in international business and transportation hubs, SIL developed a multilingual, multiaccent solution that can be run at the edge on kiosk hardware.

Using Intel® processors and the OpenVINO toolkit, SIL created Dyn-ASR, an ASR solution capable of dynamic loading and usage. This dynamic new approach allows for multiple monolingual ASR models to run at the edge without exceeding storage or memory limits. Using natively multilingual NLU, the solution can respond to end users in the language and accent they use. The flexible, configurable SIL solution allows development of contactless kiosks in new global markets and use cases.

Learn more

To discover how the SIL multilingual ASR and NLU solution can optimize language and accent recognition and enhance touchless kiosk experiences, visit ai.sil.org.

Explore Dyn-ASR capabilities in the technical paper at arxiv.org/abs/2108.02034.

About SIL International

SIL is an NGO that is collaborating with Intel to develop and deploy patent-pending, AI-driven audio technology that expands possibilities for contactless and multilingual applications in emerging markets. SIL works with local communities around the world to develop language solutions that expand possibilities for a better life. As of 2020, SIL is involved in approximately 1,350 active language projects in 104 countries. These projects impact more than 1.1 billion people within 1,600 local communities.

sil.org

Intel Distribution of OpenVINO toolkit

OpenVINO is *the* development environment for deep learning inference on Intel® hardware. It optimizes and converts models from all the major frameworks and gives developers a GStreamer-based toolset for creating inference pipelines.

Learn more ›

Intel DevCloud for the Edge

Tune, test, and deploy AI on Intel hardware, without the hardware. Intel DevCloud gives developers access to the latest Intel hardware in an online sandbox running OpenVINO in a Jupyter Notebooks dev environment. Most Intel® edge software and middleware is online and ready for testing as well. Build with Intel® components, code from scratch, or work with a point-and-click deep learning workbench.

Learn more ›

1. Internal testing and measurements by SIL. Configuration details: Intel® Core™ i9-7920X CPU, 12 core24 threads with 64 GB memory, 500 GB storage, Ubuntu 18.04. Pre-trained reference models for DeepSpeech and QuartzNet compared to optimizations with Intel® Distribution of OpenVINO™ toolkit v.2020.4. All the audio file inputs were of type 16 kHz, 16-bit PCM mono. Test by SIL as of November 2020. For more details, visit arxiv.org/abs/2108.02034.