Small Language Model (SLM) for Device AI

Akraino Robotics Blueprint, Release 8 Enhancement

Quick Background

- Akraino Robotics Blueprint
  - Led by Fujitsu and Univ Ritsumeikan, incubation 2022
  - Sponsored by SIP/Japan Cabinet Office / NEDO
  - Signalogic added real-time ASR [1]

New industries for robots


Agriculture


Food factory


Retail


Restaurant

.....

Challenges for robot in these industries

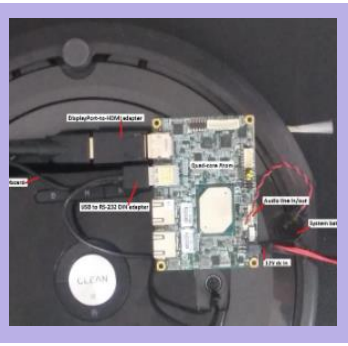
1. Objects with diverse shapes, flexibility
2. Uncertain environment (wet, clutter, customers, etc.)
3. “No cloud” communication with humans

Solutions through fusion of robot and sensors

- **Flexible robot handling using sensor data**
- **Reliable, light weight, onboard ASR [1]**

“CPS [2] Robot blueprint family” published as **OSS stack in Akraino**

PoC in progress in the field



[1] Automatic Speech Recognition

[2] CyberPhysical Systems

Device AI speech recognition challenges at the edge

- Device AI applications need to run ASR
  - With unreliable or no cloud connection
  - On very small form-factor devices (e.g. pico ITX)
  - Under difficult conditions, including background noise, urgent or stressed voice input, and background talkers
  - Robotics servo motor and other mechanical noise increases difficulty

AKRAINO FUJITSU RITSUMEIKAN Signalogic

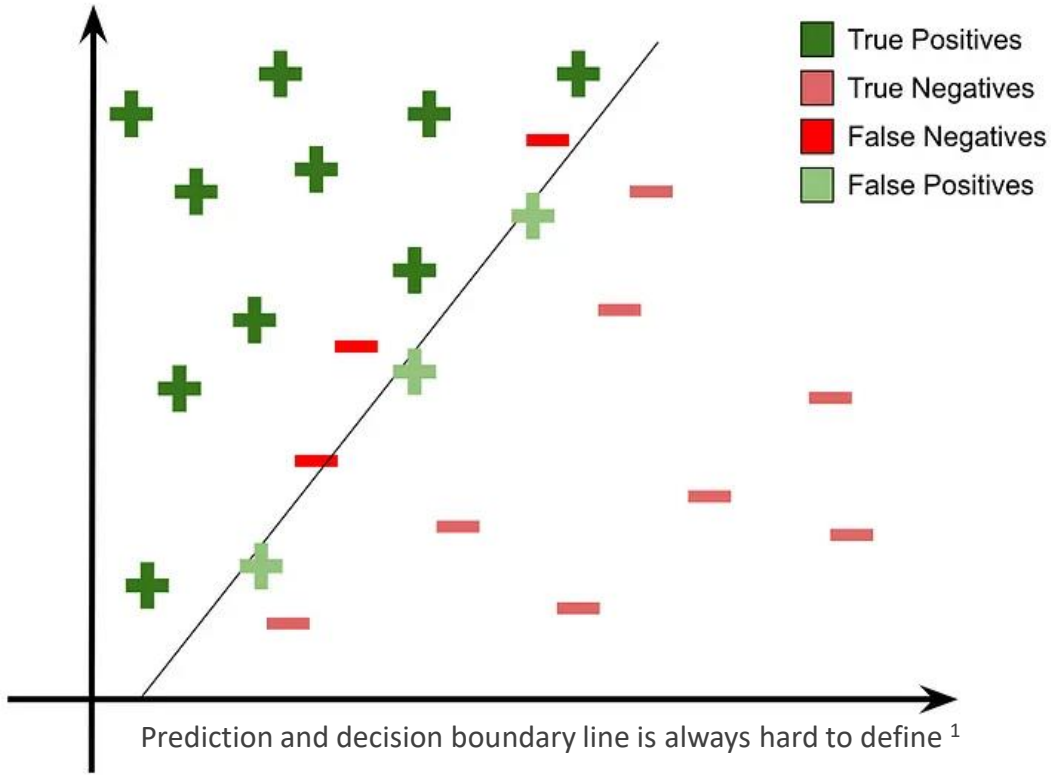
Precise Command Problem

- Machine-readable APIs must be precise
- Minimizing false positives is crucial
- Under difficult conditions, efficient open source ASRs such as Kaldi and Whisper produce "sound-alike“ errors, for example:

  “in the early days a king rolled the stake”

  which must be corrected to

  “in the early days a king ruled the state”



Prediction and decision boundary line is always hard to define [1]

- Sound-alike errors are problematic for safety and emergency situations
  - Internet / cloud connection cannot be assumed. Phones may be useless
  - A first responder may use a portable hand-held device and give commands to a robotaxi such as "get off the road in that turn-out up ahead and shut down“
  - Sometimes generalized in ASR research as “substitution errors”

AKRAINO FUJITSU RITSUMEIKAN Signalogic

[1] https://medium.com/@Sanskriti.Singh/an-emphasis-on-the-minimization-of-false-negatives-false-positives-in-binary-classification-9c22f3f9f73

Use Cases

- Factory floor personnel need to give urgent commands
  - possibly dangerous equipment (e.g. forklifts)
  - no-hands-free environments (e.g. food processing)

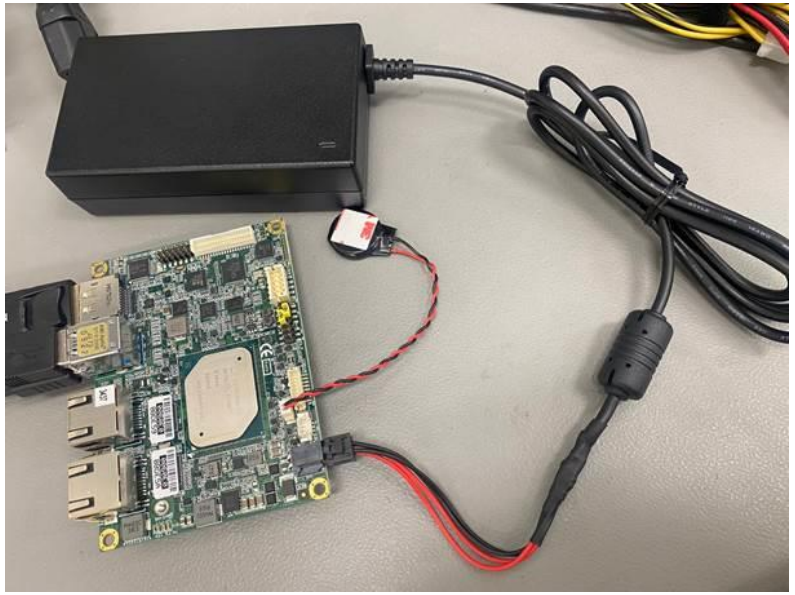
- First responders need to communicate with disabled or disconnected robotic vehicles
  - robotaxis
  - semi trucks

- Language Translation
  - sound-alike correction prior to translation
  - independent of ASR model

AKRAINO FUJITSU RITSUMEIKAN Signalogic

Requirements

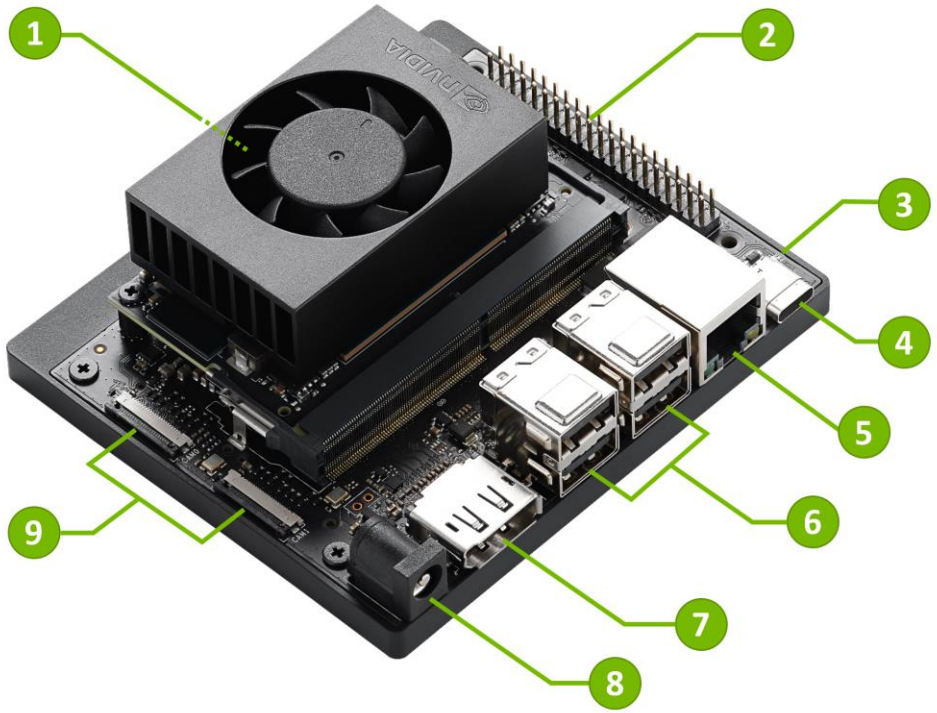
- Must correct sound-alike errors independently of ASR model without re-training, tuning, compression, or other reduction
- Very small form-factor, under 15 W
  - 4 x 3”, heat sink only, no fans
- Real-time – must run every 250 to 500 msec
  - Minimum 10 token/sec, preferably 20
- Backwards / forwards context of 5 tokens (words)
  - unlike an LLM, wide context window, domain knowledge, and extensive web page training are not needed
- Compliant with emerging teleoperation standards
  - California included teleoperation as part of its regulation for driverless vehicles in 2018
  - NIST conference in 2020
  - WiFi or USB port interfaces typical



Teleoperation and Autonomous Vehicles Overview		
	Key Information	Other Information
What is teleoperation?	• Remote operation of a machine at a distance • Requires wireless link to machine • First concepts in 1870s; wire-guided torpedoes	• Similar to remote control • Or wired link if machine is nearby • Nikola Tesla-1898: Radio-controlled boat
3 levels of AV teleoperation	• Remote monitoring of AVs • Remote assistance to AVs • Remote driving of AVs	• Monitoring of AV fleet driving • Driving assist for a short time • Driving for a substantial time
Why is it needed?	• As human backup to driverless vehicles • To manage and learn from edge cases • To gain early AV deploy with acceptable safely	• To be part of most AV regulations • Transfer edge cases to known cases • Only for specific AV use-cases
Teleoperation regulation status	• California approval granted in February 2018 • California operational use started in April 2018 • Countries: Canada, Finland, Japan, Netherlands • Shanghai and other Chinese cities	• Driverless AVs require teleoperation • AZ, FL, MI, OH, TX too; More will follow • Sweden, UK; More will follow • Teleoperation expected in China
Teleoperation use-cases	• Sidewalk AVs: Most common usage • Trucks: AV on highway; last mile teleoperation • Robotaxis: Regulation and edge case • Others: Forklifts, excavators, yard trucks, combine • Shared electric scooters	• Examples: Kiwibot, Postmates • Examples: Einride, Hub-to-hub AVs • Zoox has remote operation patent • Testing, trials, some deployment • To return to base & charging stations
Teleoperation startups	• Designated Driver: Assisted & remote driving • DriveU: Assisted & remote driving teleoperation • Ottopia: Assisted & remote-driving teleoperation • Phantom Auto: Focus on remote driving use-cases	• Teleoperation for Texas A&M shuttle • Member: Israeli teleoperation consortium • Partners: BMW, Denso, EasyMile, others • Forklifts, yard trucks and similar clients
Make or buy teleoperation?	• Top AV software platform: own teleoperation • Many companies will buy teleoperation software	• Likely integrated with AV software driver • From multiple teleoperation startups
Teleoperation standards	• Teleoperation standards likely to happen • Best chance is high level standards	• AV software driver variety is big barrier • At functional or operational level
Teleoperation Forum	• First conference on teleoperation (virtual) • NIST Vehicle Teleoperation Forum | NIST	• November 13, 2020 by NIST • 40 speakers; 8+ hours of video sessions
Teleoperation Consortium	• TC is a non-profit business organization • 30+ companies, universities, organizations	• Founded December 2020 • Website: Teleoperation Consortium
NIST=National Institute of Standards and Technology		
Source: Egil Juliussen, May 2021		

Why Not Use a GPU ?

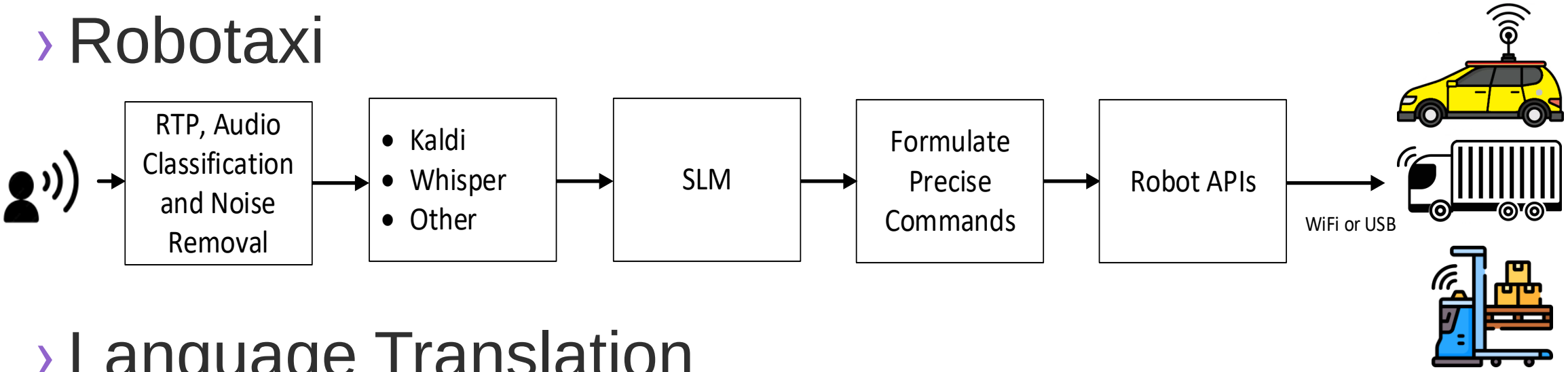
- An Nvidia Jetson Orin Nano running 7B weight Llama-2 model meets power consumption and size requirements but …
  - testing shows we can't reliably fix sound alike errors – context required too wide and model too small
  - processing rate is approx 1 token / sec
  - we still need to run RTP, audio classification, noise removal, and ASR. These are algorithms and sequential logic, extremely difficult to code in Cuda
  - we may need additional CPU cores for application-specific requirements
- Needs a fan, can't operate heat-sink only



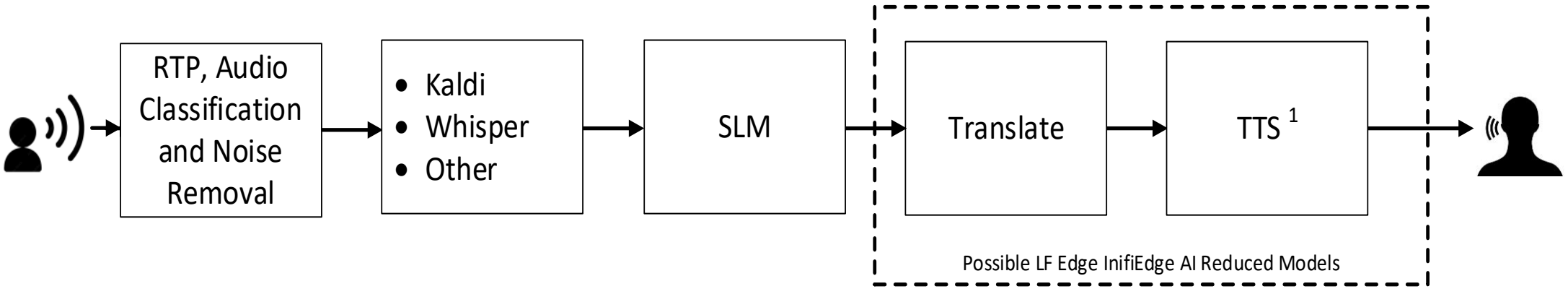
AKRAINO FUJITSU RITSUMEIKAN Signalogic

Technology Overview – Dataflow

› Robotaxi



› Language Translation



AKRAINO FUJITSU RITSUMEIKAN Signalogic

[1] Text-to-Speech

Technology Overview – Training and Inference

- Conventional CPUs
  - Arm, x86
  - no GPUs, no HBM
- Conventional memory, 8 GB min
- Training
  - frequency domain representations of 10,000 text words – becomes an image recognition problem
    - non-linear memory space, self-organizing, sound-alikes are near each other
    - extremely fast
    - no gradient descent or other high complexity algorithms
- Inference
  - content addressable memory – series of spans and local searches

AKRAINO FUJITSU RITSUMEIKAN Signalogic

Status and Next Steps

- Working now
  - RTP, audio classification, noise removal - one Atom core in real-time
  - Kaldi ASR - one Atom core in real-time
  - pico ITX board (quad core Atom x5-E3940)
  - 20,000 word vocabulary
- SLM under development
  - live demo next step
  - pico ITX board
  - planning for Akraino Fall Summit