

Small Language Model (SLM) for Device AI

Akraino Robotics Blueprint, Release 8 Enhancement



Device AI speech recognition challenges at the edge

- › Device AI applications need to run ASR¹
 - › On very small form-factor devices (e.g. pico ITX)
 - › With unreliable or no cloud connection
 - › Under difficult conditions, including background noise, urgent or stressed voice input, and background talkers
 - › Robotics servo motor and other mechanical noise increases difficulty



Precise Command Problem

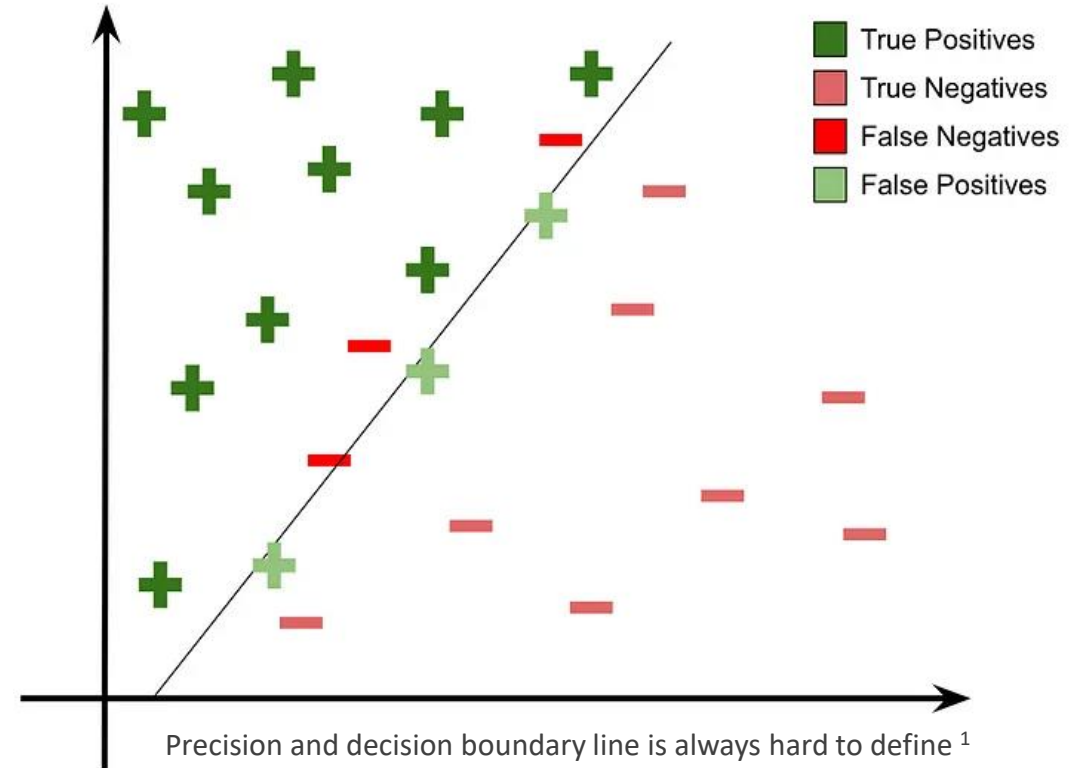
- › Machine-readable APIs must be precise
- › Minimizing false positives is crucial
- › Under difficult conditions, efficient open source ASRs such as Kaldi and Whisper produce "sound-alike" errors, for example:

`"in the early days a king rolled the stake"`

which must be corrected to

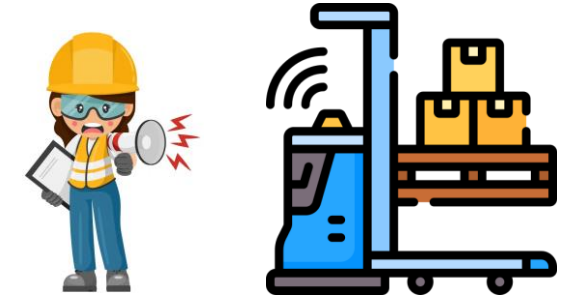
`"in the early days a king ruled the state"`

- › 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"



Use Cases

- › Factory floor personnel need to give urgent commands
 - › forklifts
 - › hands-free equipment (e.g. food processing)
- › First responders need to communicate with disabled or confused robotic vehicles
 - › robotaxis
 - › semi trucks
- › Language Translation
 - › sound-alike correction in text prior to translation
 - › independent of ASR model



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
 - › for example using two (2) Atom CPU cores
- › Real-time – must run every 300 to 500 msec
- › Backwards / forwards context of 3-4 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

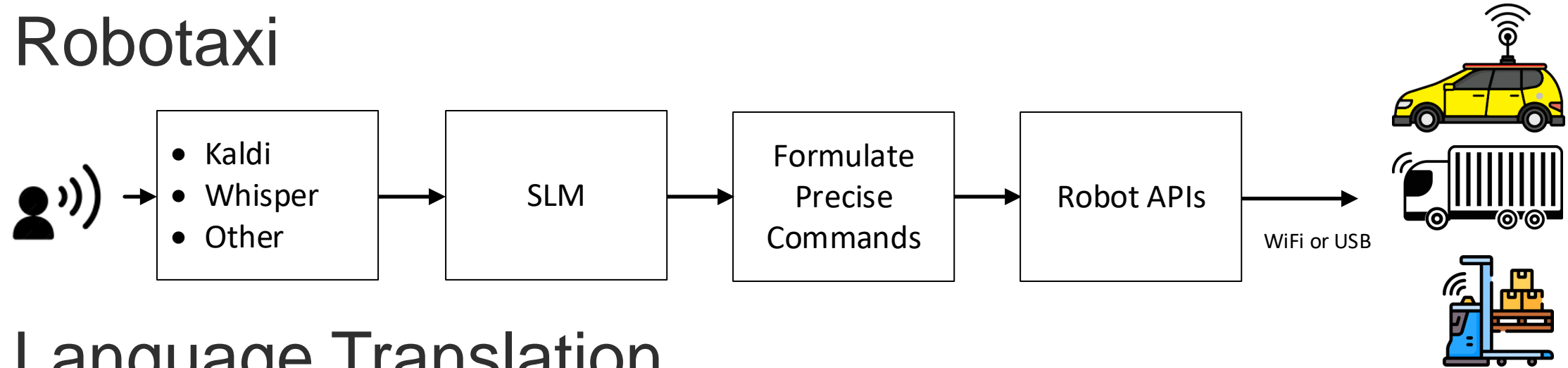


Teleoperation and Autonomous Vehicles Overview		
	Key Information	Other Information
What is teleoperation?	<ul style="list-style-type: none"> • Remote operation of a machine at a distance • Requires wireless link to machine • First concepts in 1870s; wire-guided torpedoes 	<ul style="list-style-type: none"> • Similar to remote control • Or wired link if machine is nearby • Nikola Tesla-1898: Radio-controlled boat
3 levels of AV teleoperation	<ul style="list-style-type: none"> • Remote monitoring of AVs • Remote assistance to AVs • Remote driving of AVs 	<ul style="list-style-type: none"> • Monitoring of AV fleet driving • Driving assist for a short time • Driving for a substantial time
Why is it needed?	<ul style="list-style-type: none"> • As human backup to driverless vehicles • To manage and learn from edge cases • To gain early AV deploy with acceptable safety 	<ul style="list-style-type: none"> • To be part of most AV regulations • Transfer edge cases to known cases • Only for specific AV use-cases
Teleoperation regulation status	<ul style="list-style-type: none"> • California approval granted in February 2018 • California operational use started in April 2018 • Countries: Canada, Finland, Japan, Netherlands • Shanghai and other Chinese cities 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Designated Driver: Assisted & remote driving • DriveU: Assisted & remote driving teleoperation • Ottopia: Assisted & remote-driving teleoperation • Phantom Auto: Focus on remote driving use-cases 	<ul style="list-style-type: none"> • 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?	<ul style="list-style-type: none"> • Top AV software platform: own teleoperation • Many companies will buy teleoperation software 	<ul style="list-style-type: none"> • Likely integrated with AV software driver • From multiple teleoperation startups
Teleoperation standards	<ul style="list-style-type: none"> • Teleoperation standards likely to happen • Best chance is high level standards 	<ul style="list-style-type: none"> • AV software driver variety is big barrier • At functional or operational level
Teleoperation Forum	<ul style="list-style-type: none"> • First conference on teleoperation (virtual) • NIST Vehicle Teleoperation Forum NIST 	<ul style="list-style-type: none"> • November 13, 2020 by NIST • 40 speakers; 8+ hours of video sessions
Teleoperation Consortium	<ul style="list-style-type: none"> • TC is a non-profit business organization • 30+ companies, universities, organizations 	<ul style="list-style-type: none"> • Founded December 2020 • Website: Teleoperation Consortium
NIST=National Institute of Standards and Technology		
Source: Egil Juliusen, May 2021		

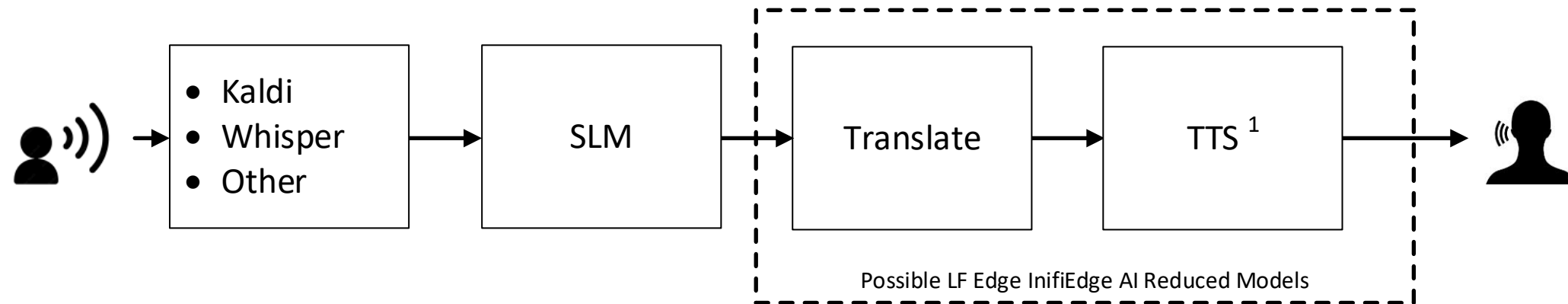


Technology Overview – Dataflow

› Robotaxi



› Language Translation



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



Status and Next Steps

- › Working now
 - › Kaldi ASR running on one Atom core in real-time
 - › pico ITX board (Atom x5-E3940)
 - › 20,000 word vocabulary
- › SLM under development
 - › live demo next step
 - › pico ITX board
 - › planning for Akraino Fall Summit

