Local commands
Contributors (in alphabetical order of last name)

Mahesh Godavarti, PhD
Engineering Technical Leader

Kelsey Kraus, PhD
Data Scientist

Hui-Ling (Vicky) Lu, PhD
Data Engineer

Melissa Samson-Fauteux, MA
Data Analysis Manager

All contributors are affiliated with the Collaboration AI team at Cisco Systems, Inc.

webex
What is Local Commands?
What is Webex Assistant?

Want to know more? Find out here.
Webex Assistant Block Diagram

User

Speech Recognition

"OK Webex" trigger word detection

Text To Speech

Webex Assistant Business Logic
Hybrid model Advantages:

- Larger model in the cloud, which can reduce an edge device’s compute capabilities
- On-device models, which are faster

Our tinyML approach builds a multilingual hybrid-model with reduced latency that mimics a natural language system running wholly at the edge.
Modeling user intents

• Intents subset queries into actions the assistant can take
• Each action facilitates an appropriate answer or task execution for the user.

1) Users use varied language to interact with an assistant (commands)
2) Model is trained on semantically similar labeled data, categorized into intents
3) Model recognizes the intent, and produces an appropriate response
Why a multilingual assistant?

Languages with at least 50 million first language speakers

- Mandarin Chinese: 11.92%
- Spanish: 5.99%
- English: 4.92%
- Hindi: 4.43%
- Bengali: 4.00%
- Portuguese: 2.87%
- Japanese: 1.66%
- Russian: 2.00%
- Other: 18.24%

Sources: Ethnologue, IMF

© 2021 Cisco and/or its affiliates. All rights reserved. Cisco Confidential
### Command selection – Variation across locales

<table>
<thead>
<tr>
<th>Linguistic considerations</th>
<th>Explanation</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Register                  | Formal vs. colloquial ways of speaking | English: Yes  
Yep | French: Oui  
Ouais |
| Verbal inflection         | How verb forms change (or not) with conjugation | English: close  
Portuguese: fechar (infinitive)  
feche (command, formal)  
fecha (command, informal) |
| Lexical variation         | Synonyms or paraphrases of the same concept | English: meeting call  
German: Meeting  
Besprechung  
Konferenz |
| Politeness level          | Formal vs. informal ways of addressing people | English: How are you?  
(How art thou?)  
Italian: (Tu) come stai?  
(Lei) come sta? |
| Gender                    | Grammatical markings indicating a noun’s gender | English: previous page  
previous call  
French: page précédente (feminine)  
appel precedent (masculine) |
| Dialectal variations      | Spoken differences based on regional standards | English:  
Call my personal room  
Brazilian Portuguese: ligar para minha sala pessoal  
Continental Portuguese: ligar para a minha sala pessoal |
| Different constructions   | Ways of expressing the same desired intent | English: Tell me a joke  
Make me laugh  
Spanish: Cuéntame un chiste  
Hazme reír |
| Syntactic variation       | Non-meaning distinctions in language use | English: Join meeting  
Join the meeting  
German: Meeting beitreten  
dem Meeting beitreten |
A closer look across locales: *decrease_volume* intent

Command sets across locales are not completely identical –

- Different counts of local commands based on language-specific variation
- Commands selected in part based on frequency of use per-language

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>decrease volume</td>
<td>baja el volumen</td>
<td>moins fort</td>
<td>leiser</td>
</tr>
<tr>
<td>lower the volume</td>
<td>bajar el volumen</td>
<td>baisse le son</td>
<td>Audio leiser</td>
</tr>
<tr>
<td>lower volume</td>
<td>bajar volumen</td>
<td>baisser le son</td>
<td>Lautsprecher leiser</td>
</tr>
<tr>
<td>reduce the volume</td>
<td>puedes bajar el volumen</td>
<td>baisser le volume</td>
<td>Ton leiser</td>
</tr>
<tr>
<td>turn down the volume</td>
<td>turn down volume</td>
<td></td>
<td></td>
</tr>
<tr>
<td>turn the volume down</td>
<td>volume down</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>Brazilian Portuguese</td>
<td>Portugal/Continental Portuguese</td>
<td>Italian</td>
</tr>
<tr>
<td>音量を下げて</td>
<td>abaixe o volume</td>
<td>baisse o volume</td>
<td>abbassa</td>
</tr>
<tr>
<td>ボリュームを下げる</td>
<td>abaixar o volume</td>
<td>baixar o volume</td>
<td>abbassare</td>
</tr>
<tr>
<td>音を小さく</td>
<td>diminua o volume</td>
<td>diminuir o volume</td>
<td>volume più basso</td>
</tr>
<tr>
<td>ボリュームを落として</td>
<td>diminuir o volume</td>
<td>reduza o volume</td>
<td>abbassa il volume</td>
</tr>
<tr>
<td></td>
<td>baixe o volume</td>
<td>reduzir o volume</td>
<td>abbassare il volume</td>
</tr>
<tr>
<td></td>
<td>baixar o volume</td>
<td></td>
<td>riduci il volume</td>
</tr>
</tbody>
</table>
# Commands on the Edge

34 intents in all languages

<table>
<thead>
<tr>
<th>Locales</th>
<th># of commands</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>186</td>
</tr>
<tr>
<td>Spanish</td>
<td>181</td>
</tr>
<tr>
<td>French</td>
<td>206</td>
</tr>
<tr>
<td>German</td>
<td>193</td>
</tr>
<tr>
<td>Japanese</td>
<td>141</td>
</tr>
<tr>
<td>Italian</td>
<td>198</td>
</tr>
<tr>
<td>Portuguese - Brazil</td>
<td>217</td>
</tr>
<tr>
<td>Portuguese - Continental</td>
<td>196</td>
</tr>
</tbody>
</table>
The model network architecture includes a common embedding network, universal to all languages, followed by a per-language decision network that captures the variations in output commands, specific to each language.

The model network is adapted from ResNet and convolutional network.
Model Training

- To reduce data collection cost, training data mostly comes from synthetic speech via various Text-to-Speech (TTS) services.
- To reduce environment mismatch between the deployment and the training condition, TTS data is augmented with real-world noise.
- The universal embedding network is trained with speech from multiple languages.
- For each specific language, the decision network is trained via transfer learning via language-specific data.
Performance Summary of Local Command Models

Performance of Command Models

Accuracy (%) vs False Alarm Rate (%)

- English
- German
- Spanish
- French
- Japanese
Local Commands Architecture Diagram

Audio Stream

Audio Pre-processing

Embedding Network

Decision Network 1

Decision m

Decision 1
Implementation

- Network developed and trained in TensorFlow
- Custom quantizer directly generates C data structures
- Fixed-point implementation in C
- Streaming Implementation
- Loop optimization
- Protobuf based model streams

<table>
<thead>
<tr>
<th>Embedding Network</th>
<th>Decision Network (Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>335KB</td>
</tr>
<tr>
<td>Model</td>
<td>1.3MB</td>
</tr>
<tr>
<td>Memory Buffers</td>
<td>700KB</td>
</tr>
<tr>
<td>Compute</td>
<td>100MMACS</td>
</tr>
<tr>
<td>Code</td>
<td>435KB</td>
</tr>
<tr>
<td>Model</td>
<td>600KB</td>
</tr>
<tr>
<td>Memory Buffers</td>
<td>110KB</td>
</tr>
<tr>
<td>Compute</td>
<td>30MMACS</td>
</tr>
</tbody>
</table>
User experience improvements

Latency (ms)

- Cloud
- Local
Conclusion

Hybrid approach can deliver good user experience by combining the best of both worlds.

Collaboration between linguistics, research, and engineering led to a “tiny”, yet effective local commands model.

• Careful selection of language-specific commands by language experts to implement on the device
• Single embedding network to be shared across multiple decision networks
• Streaming implementation of the embedding network (reusing computed variables)
• Fixed point implementation in C and loop optimization to reduce computation
What's next?

- Monitor use and user feedback
- More commands on the edge
- Higher accuracy via data
- Explore different architectures
- More languages (Italian and Portuguese)
Copyright Notice

This presentation in this publication was presented at the tinyML® Summit (March 28 - 29, 2023). The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org