Designing for
Conversational AI at the Edge

Jim Steele, VP Technology Strategy, Knowles
Intelligent Audio
Enabling Advanced Audio Solutions

Knowles creates audio solutions that differentiate our customers

OUR OFFERING
• Higher-performance microphones, advanced multi-mic integration
• DSP, smart mics, audio processors, audio software algorithms, tools, and system solutions

KNOWLES ADVANTAGE
• Cross-functional expertise from acoustics to audio algorithms
• Microphone performance leadership and scale
• Over 10B microphones shipped
### Pieces of a Conversational AI Platform

<table>
<thead>
<tr>
<th>Content Providers and Intent Execution</th>
<th>Intent Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Engines</td>
<td></td>
</tr>
<tr>
<td>Microphones and Signal Processing</td>
<td>Natural Language Understanding (NLU)</td>
</tr>
<tr>
<td></td>
<td>Automatic Speech Recognition (ASR)</td>
</tr>
<tr>
<td></td>
<td>Text to Speech (TTS)</td>
</tr>
<tr>
<td></td>
<td>Acoustic Echo Cancellation (AEC)</td>
</tr>
<tr>
<td></td>
<td>Noise Suppression</td>
</tr>
<tr>
<td></td>
<td>Continuous Voicewake</td>
</tr>
<tr>
<td></td>
<td>Always-on Voicewake</td>
</tr>
</tbody>
</table>

Example: "Book a Hotel in NYC"
"50% of all local data will be processed outside the cloud by 2022”
according to Gartner

> 42 billion connected devices
We will see amazing edge devices with new sensors & processors

Why edge?
- Low power
- Low latency
- Cost
- Privacy
- Bandwidth

From Peter Levine
Moving from Cloud to Edge: Low power

- **Going to the cloud takes power**
  - Pulling data off sensors
  - Packetizing, sending, retries
  - Waiting for response

- **Compute at the sensor**
  - Optimize for the sensor and data
  - Minimize data transfer
  - Lowest power consumption
  - Minimal hardware used

Puts a DSP+DNN compute engine right at the sensor for ultra-low power voice wake
• **Driving architecture change: real-time data**
  – Sensors (such as microphones) proliferate and provide large amounts of **real-world** data
  – Data needs to be acted on **real-time**

• **Role of Edge Intelligence**
  – Edge provides quick local **Sense-Infer-Act loop**
  – Cloud provides longer global **Learning loop**

• **Network Reliability**
  – Can add unacceptable latency or fail outright
Moving from Cloud to Edge: Cost

Cloud Speech-to-Text API pricing

Powerful speech recognition.

Cloud Speech-to-Text is priced per 15 seconds of audio processed after a 60-minute free tier. For details, please see our pricing guide.

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>0-60 MINUTES</th>
<th>OVER 60 MINUTES, UP TO 1 MILLION MINUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Recognition (all models except video)</td>
<td>Free</td>
<td>$0.006 USD / 15 seconds*</td>
</tr>
<tr>
<td>Video Speech Recognition</td>
<td>$0.006</td>
<td>$0.012 USD / 15 seconds*</td>
</tr>
</tbody>
</table>

This pricing is for applications on personal systems (e.g., phones, tablets, laptops, desktops). Please contact us for approval and pricing to use the Speech-to-Text API on embedded devices (e.g., cars, TVs, appliances, or speakers).
Moving from Cloud to Edge: Privacy

- **Differential Privacy**
  - Anonymous data sent to cloud for crowdsourced learning

- **On-device Intelligence for Personalization**
  - Machine learning at the edge required

“We try to keep as much of your information on that device as possible, because we want the device to ‘know’...because you count on the device to be smart for you.”

-- Tim Cook, CEO, Apple
Moving from Cloud to Edge: Bandwidth

- **Advent of Machine Learning**
  - Aggregate as much data as possible from as many sensors as possible to learn “the truth”
  - Deep learning inferences provide better accuracy
  - Continuous training for accuracy and automation

- **Role of hardware**
  - Consume and process as much data as possible as close to the sensors as possible
  - Plethora of use-cases dictates need for OpenDSP with machine learning accelerators

Even the RAM industry needs new technology to handle the increasing bandwidth required for machine learning. HBM – High Bandwidth Memory
## Machine Learning Compute on the Cloud vs. the Edge

<table>
<thead>
<tr>
<th></th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Edge</strong></td>
<td>• Low latency</td>
<td>• Compute memory may be limited</td>
</tr>
<tr>
<td></td>
<td>• Can consume lots of data from many sensors</td>
<td>• Limited access to latest neural network model</td>
</tr>
<tr>
<td></td>
<td>• Easier to maintain privacy</td>
<td>• Algorithms not always accessible to developers</td>
</tr>
<tr>
<td></td>
<td>• No Internet connection needed</td>
<td></td>
</tr>
<tr>
<td><strong>Cloud</strong></td>
<td>• Always learning</td>
<td>• High latency</td>
</tr>
<tr>
<td></td>
<td>• Nearly unlimited compute and memory</td>
<td>• Continual cost</td>
</tr>
<tr>
<td></td>
<td>• Algorithms as a service ($$$)</td>
<td>• Privacy and security issues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• High/continuous bandwidth to the cloud may be required</td>
</tr>
</tbody>
</table>
Deep Learning Chipsets are becoming more prevalent

Tractica expects that, in the long term, almost every chip will have some sort of AI capability embedded in it.
Enabling Advanced Audio Solutions

Knowles creates audio solutions that differentiate our customers

OUR OFFERING
• Higher-performance microphones, advanced multi-mic integration
• DSP, smart mics, audio processors, audio software algorithms, tools, and system solutions

KNOWLES ADVANTAGE
• Cross-functional expertise from acoustics to audio algorithms
• Microphone performance leadership and scale
• Over 10B microphones shipped