Sensory's TrulyHandsfree and Arm's Cortex-M55: Achieving High Wake Word Accuracy with Less Resources for Low-power, Customizable Voice Control

Case Study

Goal

Efficient wake word recognition on microcontrollers with Cortex-M55 and Helium technology for use in consumer and automotive products that include more and more AI features for voice applications.

Challenge

Companies want to create their own branded voice experiences and strengthen their relationship with their customers. When seeking to implement a custom, branded wake word, there is always a compromise between the desired accuracy and the resources required to achieve it. On a constrained device, this is a particular challenge because the memory and processing resources are limited by the available hardware. Also, when designing for battery-powered devices, energy efficiency is a primary concern for always listening wake word implementations. Customers expect performance equal to or better than established smartphone and smart speaker experiences. The challenge is to achieve the best possible accuracy within the constraints of the platform.

Solution

Sensory's TrulyHandsfree: Always-Listening Embedded Speech Recognition

Sensory's wake word and phrase spotting technology is known for fast response, low power consumption, and excellent performance from a distance or in noisy environments. This technology is an integral component for fully featured voice control of devices in the home, car, and anywhere voice user interfaces could be deployed.

The combination of Sensory's optimized software with the performance of the Arm Cortex-M55 processor is a compelling solution.

Benefits

- Best in class performance (false reject, false accept) for an optimal user experience.
- Supports dozens of languages and enables global voice coverage.
- Flexible, model sizes from 1MB to as small as 40KB, which can be customized for DSP, microcontroller, or applications processor-based products.

Applications

Battery operated devices such as True Wireless Headsets, TV Remotes and Smart Watches

AC mains powered devices such as Smart Speakers, Smart TVs and Smart Appliances

Automotive Infotainment Systems and Mobile Applications

Learn more about the products we love at Sensory: https://www.sensory.com/featured-products/.

Design challenges

Microcontrollers are typically designed for resource-constrained applications. As more and more features and functions are squeezed into products that utilize AI on the endpoint, Sensory's engineers are constantly challenged to do more, with less. For TrulyHandsfree, this translates into maintaining optimal wake word accuracy with a reduction in MHz.

To provide a successful voice experience and unlock a whole new range of applications where ML/AI at the endpoint is pushed even further, our team needed to explore the possibilities with a microcontroller capable of handling DSP-type workloads in an efficient way. We needed to investigate and quantify the benefits that could be directly applied to keyword spotting products such as true wireless earbuds, wearable health trackers, smart speakers and video doorbells. Products which all need to stay competitive, by packing in more and more AI features.

Design implementation

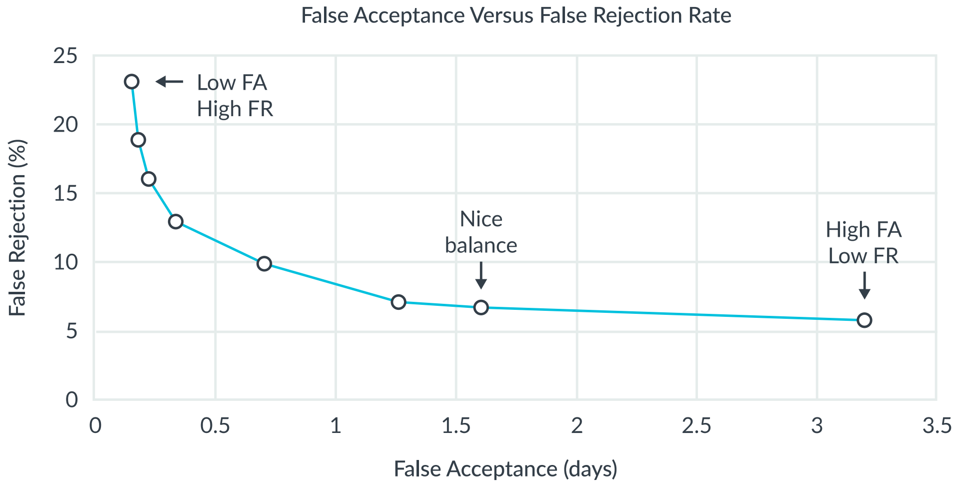
Arm Cortex-M55 offered a solution that would enable Sensory to bring AI to more devices and people in the most efficient way. As we have been working on Cortex-M4 and have had a long history working with Arm's toolchain, our engineering team jumped on the opportunity to port our existing TrulyHandsfree software to Cortex-M55.

The Cortex-M55 processor brings more performance, simplified software development, and an extensive ecosystem, which enabled Sensory to conduct investigations in an efficient way and enabled us to quantify the benefits of Cortex-M55 for the keyword spotting applications.

Before we jump straight into the Cortex-M55 benefits, it is probably best to revisit a few high-level topics on how to measure wake word accuracy and how model size and MHz come into play.

In general terms, wake word accuracy is represented as an operating point on the performance graph. The operating point is meant to provide a balance between false reject (FR) percentage and false accepts (FA) per day. You can see this on the chart below. Also, a detailed report on FR/FA testing can be downloaded here.

Figure 1
False acceptance vs.
false rejection rate



There are many different factors when defining the actual operating point, but in general the balanced approach is desired.

The recurring theme of doing more with less, for AI at the endpoint presents the greatest challenge and often means that efficiency or accuracy has to be sacrificed. For example, if developers are limited in processing power and require fewer MHz for wake word detection, then the typical solution would be a smaller wake word model which results in lower frequency, but also lower accuracy.

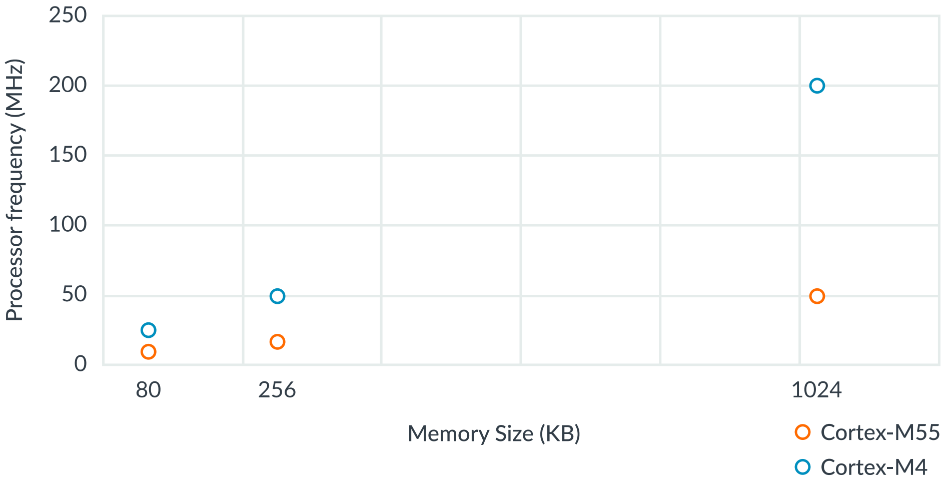
This is not ideal, but sometimes a design trade-off that has to be made. The table below shows an estimate on how model size, MHz, and False Reject rate on a Cortex-M4, which we have been using in our current solution. For comparison purposes, the wake word engine is being exercised in a variety of background noise profiles and each False Reject rate is referenced to a fixed False Accept rate of three events per day.

TrulyHandsfree Model Size	MHz Cortex-M4	False Reject % Lower is Better*
1MB	~200	2-5%
250KB	~50	5-10%
80KB	~20	10-20%
*Estimates		

With each reduction in model size there is a significant reduction in MHz, which is also accompanied with a loss of accuracy. A loss of accuracy that creates the potential for a higher occurrence both False Rejects (FR) and/or False Accepts (FA). However, things are now different with Cortex-M55.

Arm's new Cortex-M55 includes Arm Helium technology, a new vector instruction set extension that provides a significant uplift when applied to Sensory's TrulyHandsfree. When compared to Cortex-M4, which is currently used in TrulyHandsfree, Cortex-M55 provides equivalent wake word accuracy, but with an average reduction in MHz of 73%.

Figure 2
Cortex-M4 vs. Cortex-M55
MHz by model size



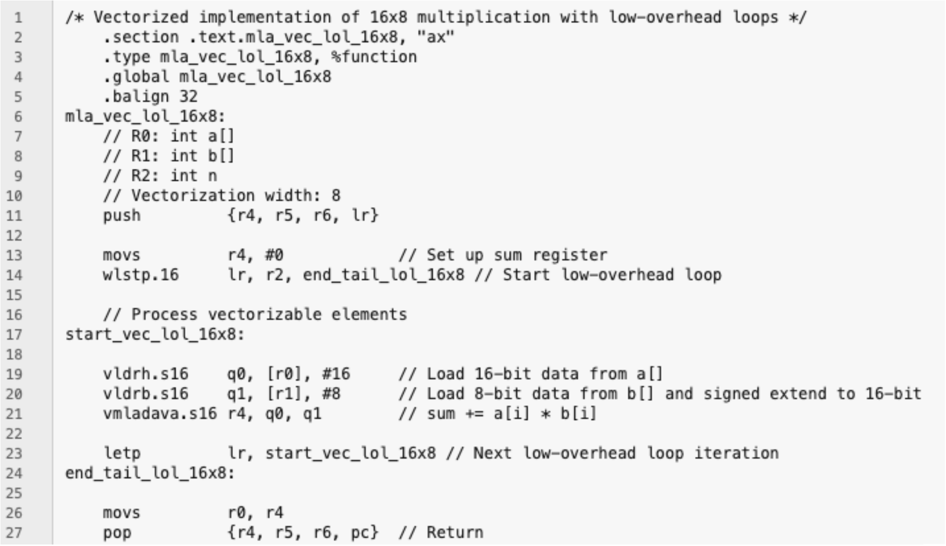
This is an initial analysis, Sensory expects that even further optimization may be achieved with Cortex-M55.

Working with Sensory's TrulyHandsfree and Arm's Cortex-M55, developers no longer face the trade-off of decreased accuracy for lower frequency. When paired together, the solution enables developers to free up resources and simultaneously maintain model size for industry leading, wake word accuracy. With such a dramatic reduction in MHz, some developers may even choose a more accurate model and still leverage a lower frequency. From a software perspective, developers can take advantage of the simplified experience working with a single toolchain and a familiar software development ecosystem. Cortex-M55 is an efficient solution for workloads for which you would often use a DSP, but now you do it with only one processor and a familiar toolchain.

Sensory and Cortex-M55 products provide a premium user experience for wake word spotting and, with a lower frequency, they open up the resources for powerful new features like Sound Identification or gesture control. Arm and Sensory empower developers to do more with less. Saving compute resources leads directly to greater energy efficiency and enables new use cases. For many constrained use-cases, energy usage is of critical importance. The inclusion of Helium vector processing technology in the Cortex-M55 processor enables significant improvements in terms of useful work per Joule when running DSP and ML workloads. Simulations carried out by Arm show an improvement of over x3 in average energy efficiency compared to Cortex-M4, measured over a range of common DSP kernels.

To see how the features available in the Cortex-M55 processor allowed us to optimize the implementation of Truly HandsFree, let's look at the following example.

The code example below shows the use of low overhead loops and vectorized instructions to achieve a smaller model size and much higher performance.



```
/* Vectorized implementation of 16x8 multiplication with low-overhead loops */
    .section .text.mla_vec_lol_16x8, "ax"
    .type mla_vec_lol_16x8, %function
    .global mla_vec_lol_16x8
    .balign 32
mla_vec_lol_16x8:
    // R0: int a[]
    // R1: int b[]
    // R2: int n
    // Vectorization width: 8
    push         {r4, r5, r6, lr}

    movs         r4, #0            // Set up sum register
    wlstp.16     lr, r2, end_tail_lol_16x8 // Start low-overhead loop

    // Process vectorizable elements
start_vec_lol_16x8:

    vldrh.s16    q0, [r0], #16     // Load 16-bit data from a[]
    vldrb.s16    q1, [r1], #8      // Load 8-bit data from b[] and signed extend to 16-bit
    vmladava.s16 r4, q0, q1        // sum += a[i] * b[i]

    letp         lr, start_vec_lol_16x8 // Next low-overhead loop iteration
end_tail_lol_16x8:

    movs         r0, r4
    pop          {r4, r5, r6, pc}  // Return
```

We can also see above that native support for 8-bit datatypes in the Armv8.1-M architecture implemented by Cortex-M55 allows for significant optimizations. There are many similar examples where we have been able to exploit the capabilities offered by Helium. Collectively, these gains lead to very significant improvements in processing efficiency, which directly drive system cost and energy savings.

Why Arm

- When Sensory's TrulyHandsfree is combined with Arm's Helium vector processing, we achieve best in class performance with a substantial reduction in cycles. Customers can expect the same accuracy with a reduction in MHz of 25% to 75%.
- The Arm ecosystem and the developer community brings the support needed to develop our solutions in the most efficient way.

Looking ahead

Sensory sees a stimulating future where multiple voice assistants will reside on the same device. End users will be able to speak directly to the brands and companies that are important to them. Each of these interactions will be supported by a custom branded wake word. Conversations with technology will flow seamlessly across mobile applications, smart speakers, cars and wearables. Sensory is excited to be a part of this voice technology evolution.

About Sensory

Sensory Inc. creates a safer and superior UX through vision and voice technologies. Sensory's technologies are widely deployed in consumer electronics applications including mobile phones, automotive, wearables, IoT and various mobile apps. Sensory's product line includes TrulyHandsfree voice control, TrulySecure biometric authentication, and TrulyNatural large vocabulary natural language embedded speech recognition.

Company website: https://www.sensory.com/wake-word/
TrulyHandsfree product brief: https://info.sensory.com/en-us/thf-brief.

Sensory is a member of the Arm AI partner program.