WHITEPAPER

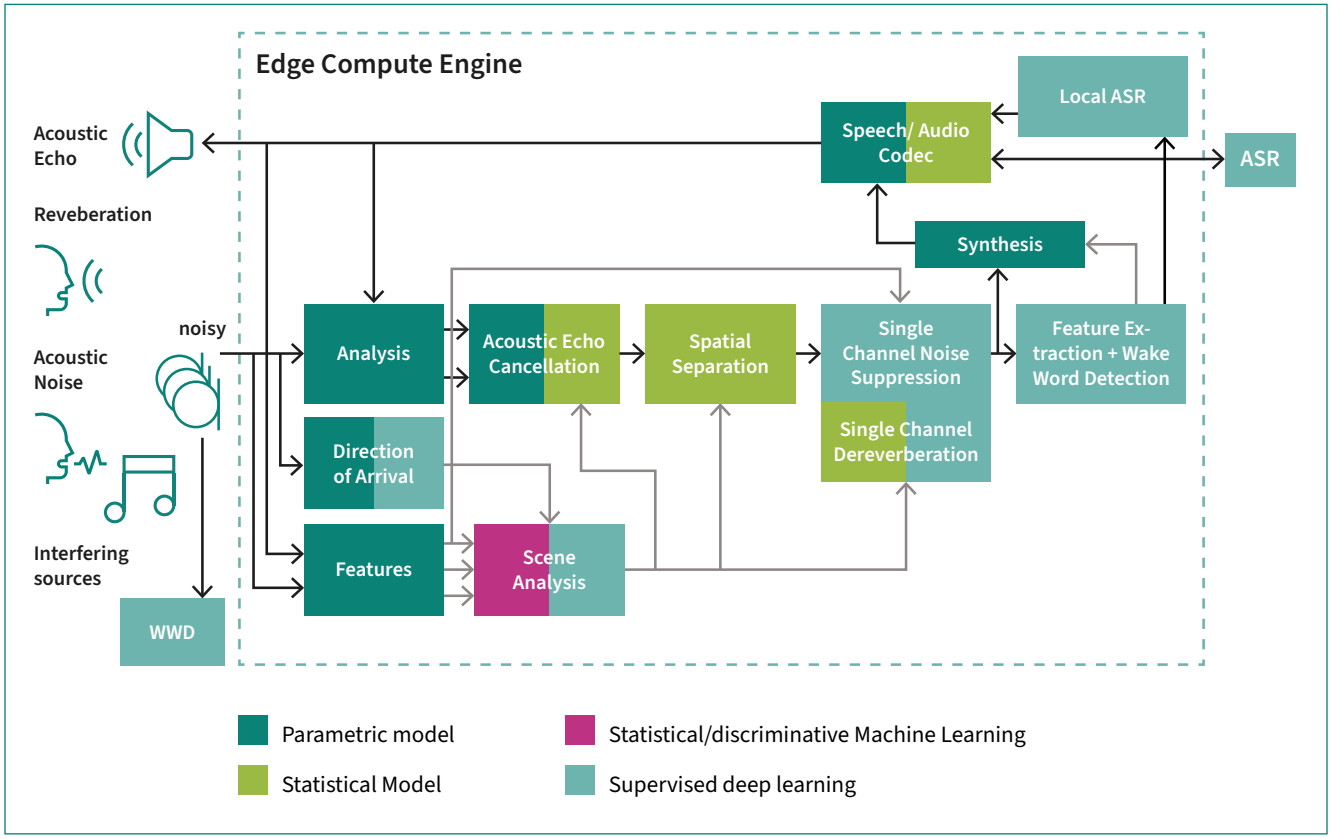
Power Efficient Audio and Voice Solution on PSOC™ Edge E84

www.infineon.com

Abstract

Voice based smart assistants are becoming increasingly efficient and accurate. They have become ubiquitous and makes user's life more convenient, productive, and efficient. Integrating this technology in more devices would increase their usefulness. Devices like wearables, smart home devices, speakers, cars etc. can benefit from integrating this technology.

Adding voice to your application comes with its own set of challenges. You need to perform a lot of signal processing, on the audio that you receive from microphones, such as Acoustic Echo Cancellation (AEC), beamforming, noise suppression, de-reverberation, etc. For faster response times you need to run wake-up word detection and speech recognition locally on the device. To do this you need to run machine learning models on the device instead of sending them over to the cloud for processing. Conventionally this is done on MPU-dominated Cortex®-A5 or Cortex®-A7 cores with higher processing capability but high-power consumption, cost and complexity. This limits the functionality on some embedded devices like smart watches which require low power consumption and lower cost. A typical audio and voice application looks like the following diagram.



As you can see, voice and audio applications requires different kinds of data processing, some of which requires vector processing while others require a neural net accelerator.

PSOC™ Edge provides a compelling solution to integrate voice and audio on battery powered, low-cost embedded devices. With a Cortex®-M55 core coupled with Helium™ DSP extensions and Ethos™-U55 microNPU coprocessor, PSOC™ Edge has enough compute power to meet the local voice processing workloads. With an efficient low power analog subsystem, a light-weight neural network accelerator and Cortex®-M33 core one can enable always-on wake-up word detection on battery powered devices without compromising battery life.

This whitepaper describes the various building blocks available on PSOC™ Edge device which can enable low-power, always-on, on-device audio and voice applications.

Table of contents

1 PSoC™ Edge E84 MCU

PSOC™ Edge E84 MCU product line is a low-power with high performance MCU family, designed for compute performance, human-machine interface (HMI), machine learning (ML), enhanced sensing, real-time control, and low-power applications.

This product line is a dual-CPU microcontroller with a neural net companion processor, DSP capability and high-performance memory expansion capability (QSPI). It comes with a slew of peripherals like low-power analog subsystem with high-performance analog-to-digital conversion and low-power comparators, IoT connectivity, communication channels, and programmable analog and digital blocks. PSOC™ Edge boasts seamless integration with 2.5D GPUs, to enable a rich graphical user interface.

ModusToolbox™ software is a modern, extensible development environment supporting a wide range of Infineon microcontrollers, including PSoC™ Arm® Cortex® microcontrollers, and various Infineon connectivity options. ModusToolbox™ development environment includes installable SDKs and libraries; industry-standard Arm® tools; RTOS support, robust and easy-to-use ML and HMI software and tools. Functions supported include security, communications and control, and DSP capability, in a multi-domain architecture which enables fine-grained power optimization and dynamic frequency and voltage scaling.

The always-on domain of the PSOC™ Edge E84 supports voice recognition, wake-on-touch, battery monitoring, and other sensing applications. These functions are provided at extremely low power.

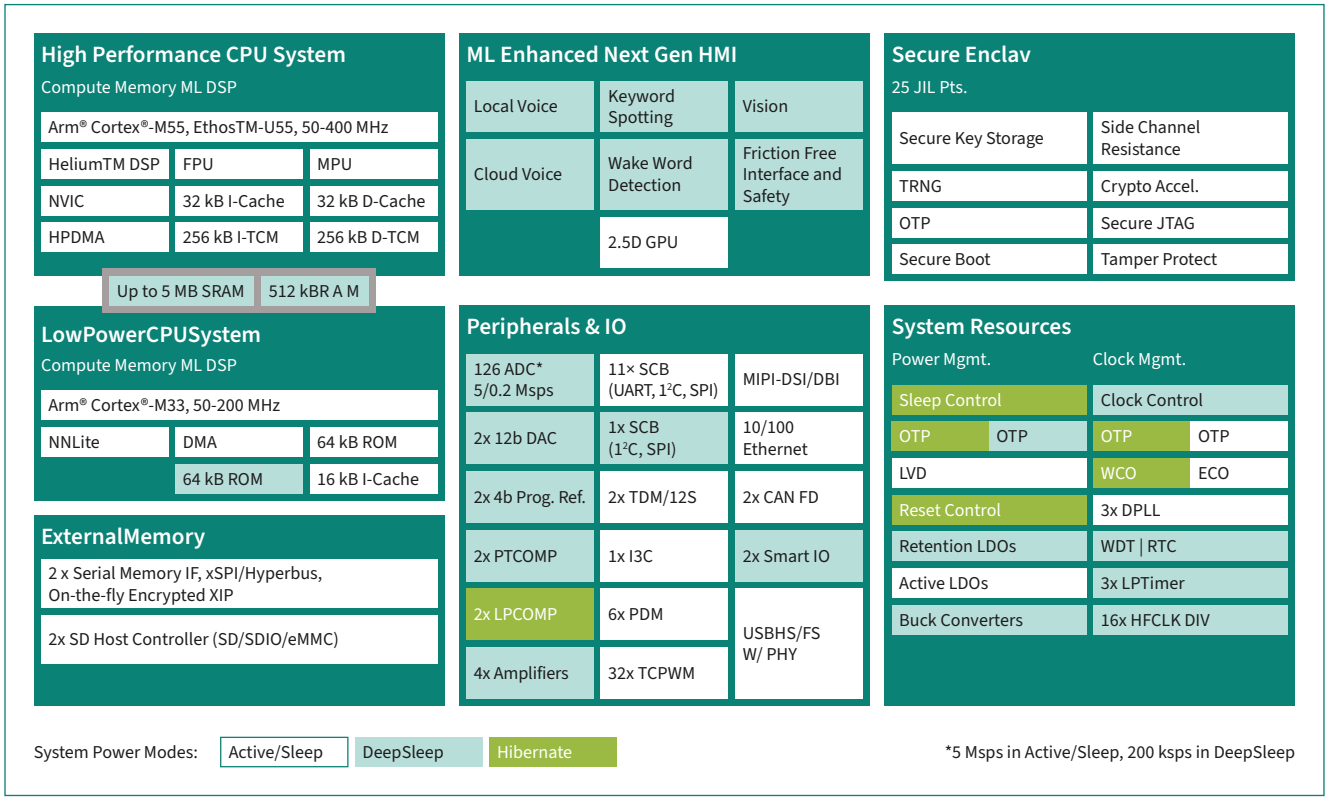


Figure 1 PSOC™ Edge E84 Block Diagram

2 PSOC™ Edge E84 Audio and Voice Features

2.1 Audio and Voice Architecture

The following diagrams shows the various software and hardware blocks available in PSOC™ Edge for audio and voice.

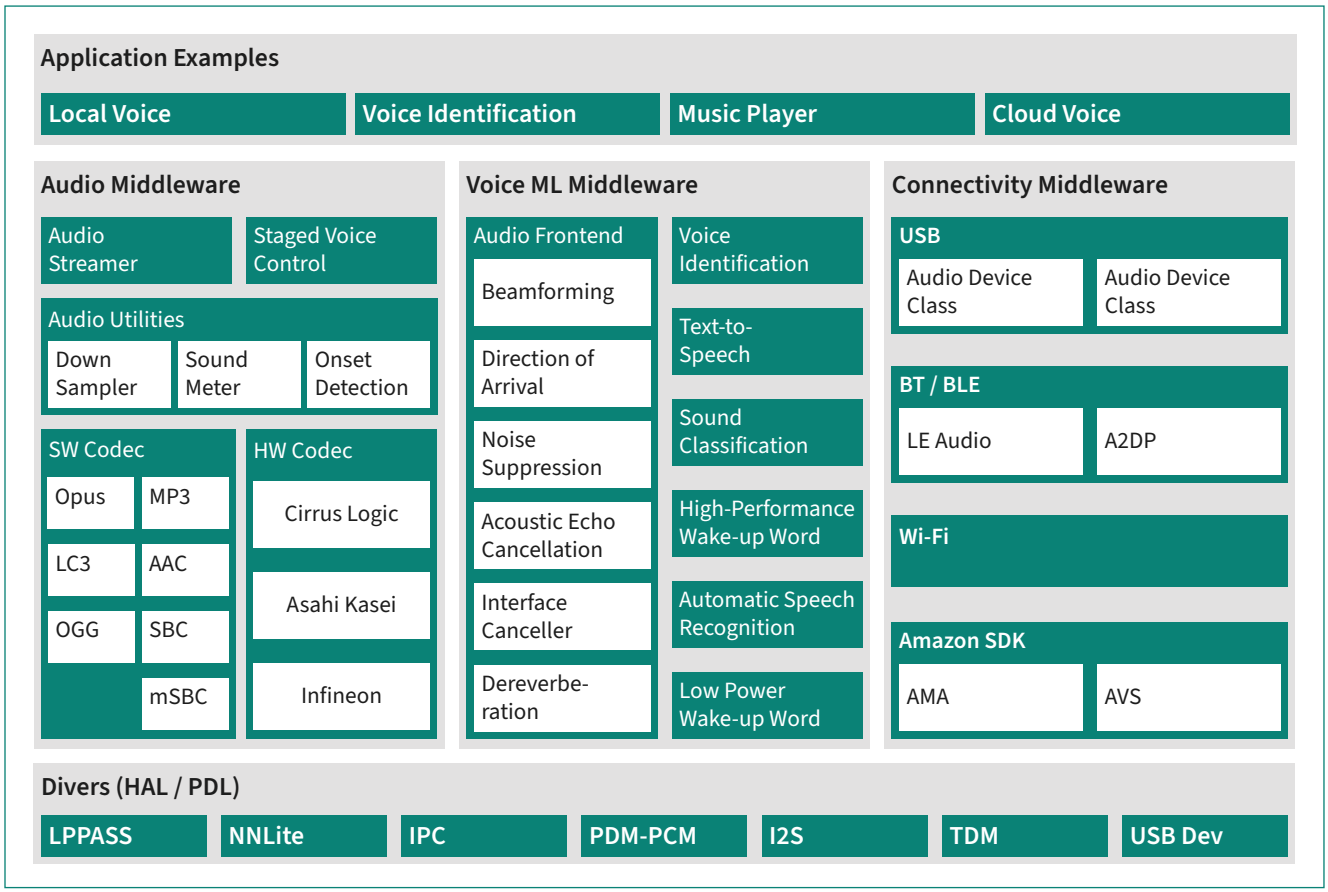


Figure 2 PSOC™ Edge Audio and Voice Architecture

Audio use cases require a combination of hardware and software blocks to work in sync for the best results. The hardware blocks are used to detect acoustic activity, convert the analog audio signals to digital audio signals, buffer the data as well as for playback. The software blocks are required to process the audio like removing noise, echo cancellation, speech recognition etc. Infineon and its partners will provide all these building blocks to simplify customer's audio and voice applications development. In the following sections we will go into more details.

2.2 Hardware Blocks

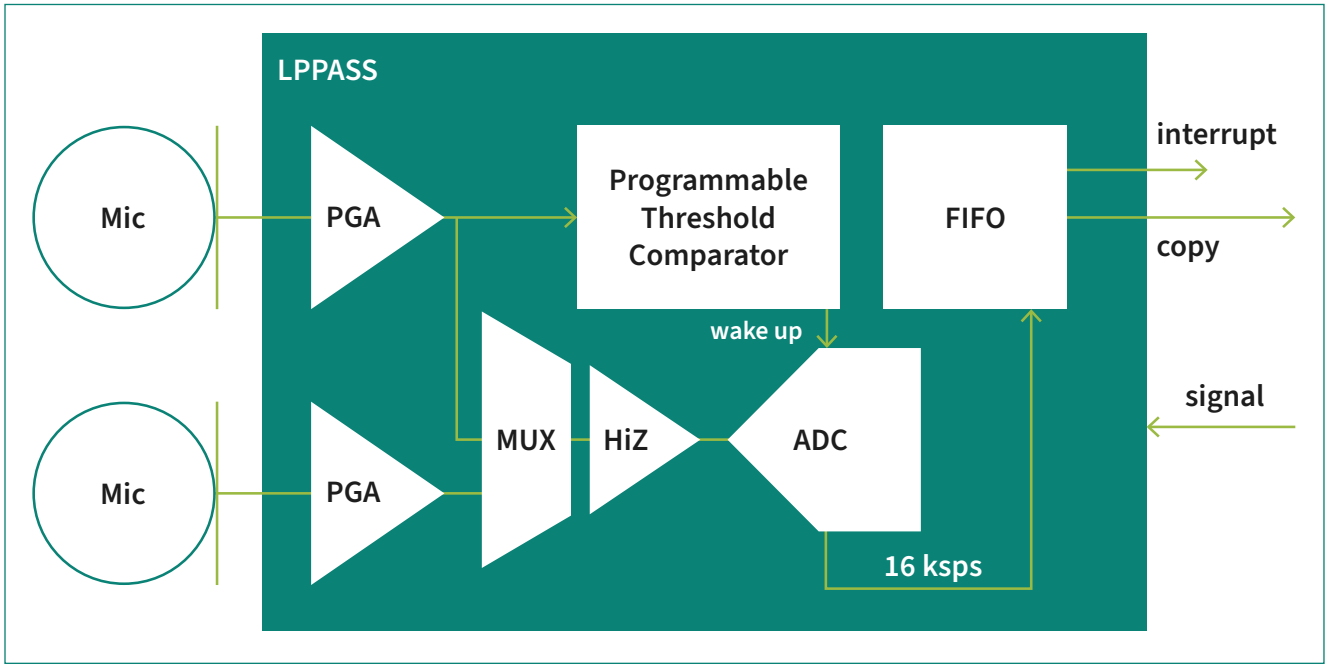
Enabling low-power always-on voice functionality on your product requires specialized hardware blocks which can run continuously on extremely low power to detect any audio activity and identify the wakeup word. PSOC™ Edge provides a series of such hardware blocks.

2.2.1 Audio Input/Output

PSOC™ Edge provides multiple options to record and playback audio which enables flexibility for the developer to optimize the system for his specific use case. The audio input output blocks of the PSOC™ Edge device can be controlled by either the Cortex®-M33 or Cortex®-M55 core and in some cases can work independently thus allowing these cores to be in low power mode.

– Analog Mic

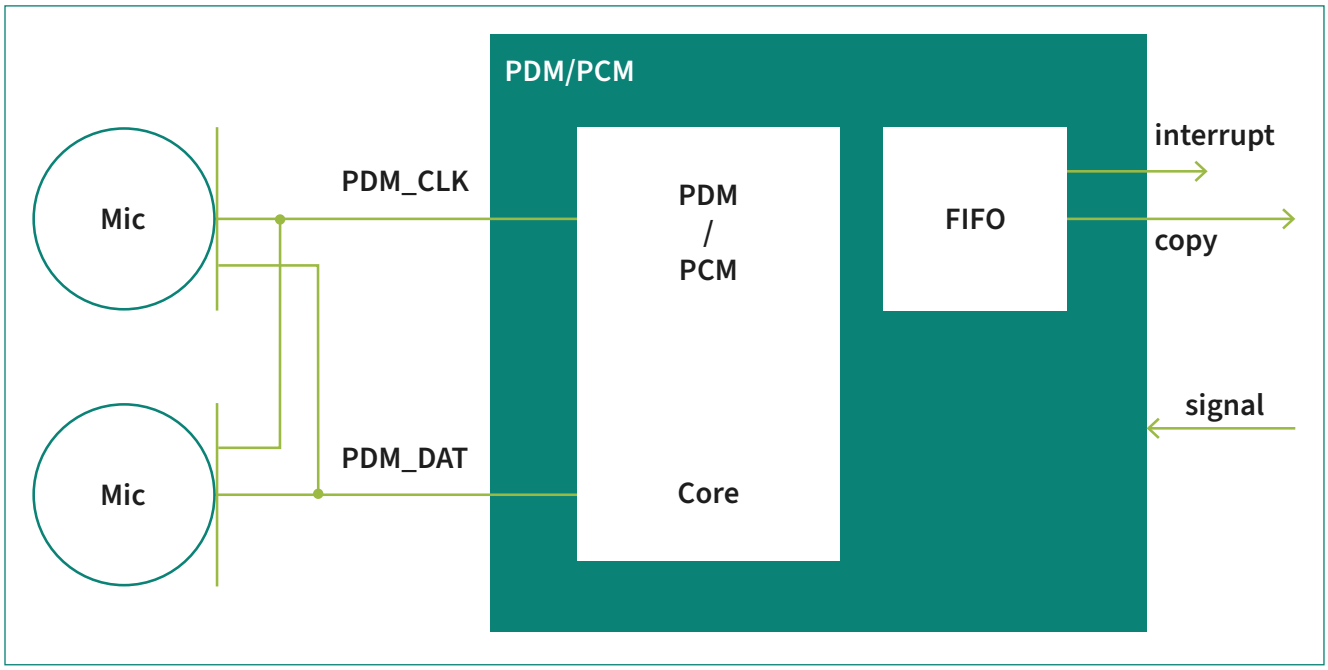
The Low Power Programmable Analog Sub System (LPPASS) block in the PSOC™ Edge device can connect up to two analog microphones and convert the analog audio signals to digital signals for further processing.



The LPPASS block can function in system deep sleep mode where rest of the device, including the Cortex®-M33 and Cortex®-M55 cores, are off. This enables always-on voice detection without draining the battery. One of the mics can be connected to a Programmable Threshold Comparator (PTC). This comparator allows detection of audio above a certain threshold. Once a loud enough audio is detected it can trigger the Analog to Digital Converter (ADC) to sample both the mics. This further reduces power consumption by limiting the use of ADC. There are other components such as Programmable Gain Amplifiers (PGA), Multiplexers, FIFO buffer etc which allows smooth operation of this system.

– Digital Mic

The Pulse Density Modulation/Pulse Code Modulation (PDM/PCM) block supports up to two digital microphones.



Using the digital microphones can result in higher power consumption for always-on use cases but can result in lower system complexity as the developer doesn't have to configure the analog front-end. The data is stored in the FIFO buffer and an interrupt is generated once there is enough data.

– Audio Output

PSOC™ Edge supports two options for audio output: The I2S IP block in PSOC™ Edge can be connected to an external hardware codec which is in turn connected to a loudspeaker. The I2S block can also be configured for the TDM interface which is useful for sending multi-channel data. This same output data can also be passed to the Audio Echo Cancellation (AEC) block for better wakeup word detection. The PSOC™ Edge also supports the Bluetooth host stack which can send audio data over Bluetooth links to a speaker or earbuds.

2.2.2 Machine Learning

Machine learning applications are essential for voice applications. They enable classification of audio into various categories as well as detect certain patterns like a specific keyword or entire queries. Specialized hardware blocks are required to run these machine learning models in a power efficient manner. PSOC™ Edge provides the following hardware blocks to run these sophisticated models.

– NNLite

NNLite is a low-power neural processing unit that accelerates specific set of Neural Network (NN) inference calculations optimized for low to moderate complexity Machine Learning (ML) models such as Wake Word Detection (WWD) and Human Activity Recognition (HAR) ideal for always-on applications and use cases. This block can execute lighter neural network workloads while consuming lower power.

– Arm Cortex®-M55 with Helium™ Extension

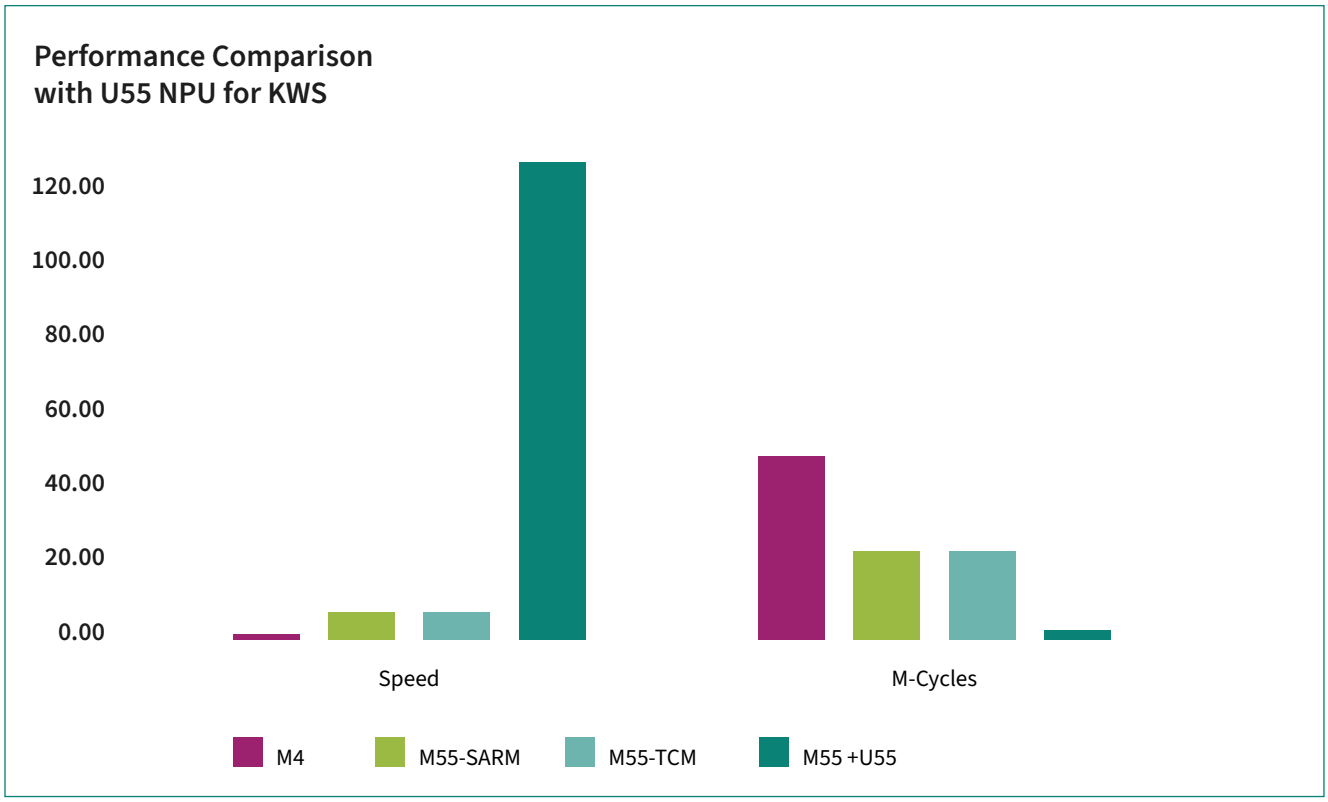
The Arm Cortex®-M55 processor is Arm's most AI-capable Cortex®-M processor and the first to feature Arm Helium™ vector processing technology. Helium™ technology is a vector extension designed for low-power embedded Systems. The following table gives the AudioMark benchmark numbers with Helium™ extension disabled and enabled.

Audiomark benchmarking		
	Helium™ OFF (Cycles	Helium™ acceleration (Multiple)
Beamforming	461821	3.14
AEC	811924	1.98
ANR	1449389	3.25
MFCC	140448	3.66
Inference	8571891	3.91
Average Helium™ Acceleration		**3.19**

The Arm Cortex®-M55 + Helium™ extension can provide up to 3 times better performance.

– Ethos™-U55 NPU

Ethos™-U55 is a first generation microNPU that can work with a Cortex®-M processor. It allows acceleration of neural networks in an extremely low-area and with low-power consumption compared to a MCU or MPU based execution. While Cortex®-M55 with its built in Helium™vector processing extensions can run ML models on the embedded micro-controller platforms, integration with the Ethos™-U55 microNPU delivers better ML performance compared to previous generations of Cortex®-M processor family. U55 with the same software stack can increase the ML performance of a Cortex®-M55 system by up to 30x. The following chart shows the relative performance of Cortex®-M4, Cortex®-M55 with and without tightly coupled memories (TCM) and Cortex®-M55 + U55 for a keyword spotting application.



From the chart we can see that Cortex®-M55 with its Helium™ extensions offers slightly better performance for ML uses cases compared to Cortex®-M4 MCU. However, combining Cortex®-M55 with Ethos™-U55 micro NPU enhances the system performance for various ML use cases making it possible to develop complicated ML applications on embedded platforms.

2.2.3 Software Blocks

– Audio Front-End (AFE)

The Audio Front-End middleware process the incoming audio to improve its quality. It can pick up the user's voice, amplify it and remove background noise by running various algorithms such as:

Beamforming

Beamforming technique uses two or more microphones to form a spatial filter which can extract a signal from a specific direction and reduces the contamination of signals from other directions.

Direction of Arrival

Direction of Arrival estimates the relative position of the sound with respect to the microphone.

Noise suppression

Noise suppression reduces stationary background noise.

Acoustic Echo Cancellation

Acoustic cancellation technique cancels acoustic feedback between a speaker and a microphone in loud-speaking audio systems.

Interference Canceller

An interference canceller uses a sample of the interfering signal to generate a real-time anti-interference signal that is the exact opposite of the interfering signal. The interference canceller combines the interference and anti-interference signals, cancelling each other out.

Dereverberation

As the name suggests dereverberation removes the reverberation from sound.

This middleware is assisted by a GUI based configurator/tuner which can be used by the developer to set parameter for the various algorithms he wants to run and tune their system for mic gain and AEC.

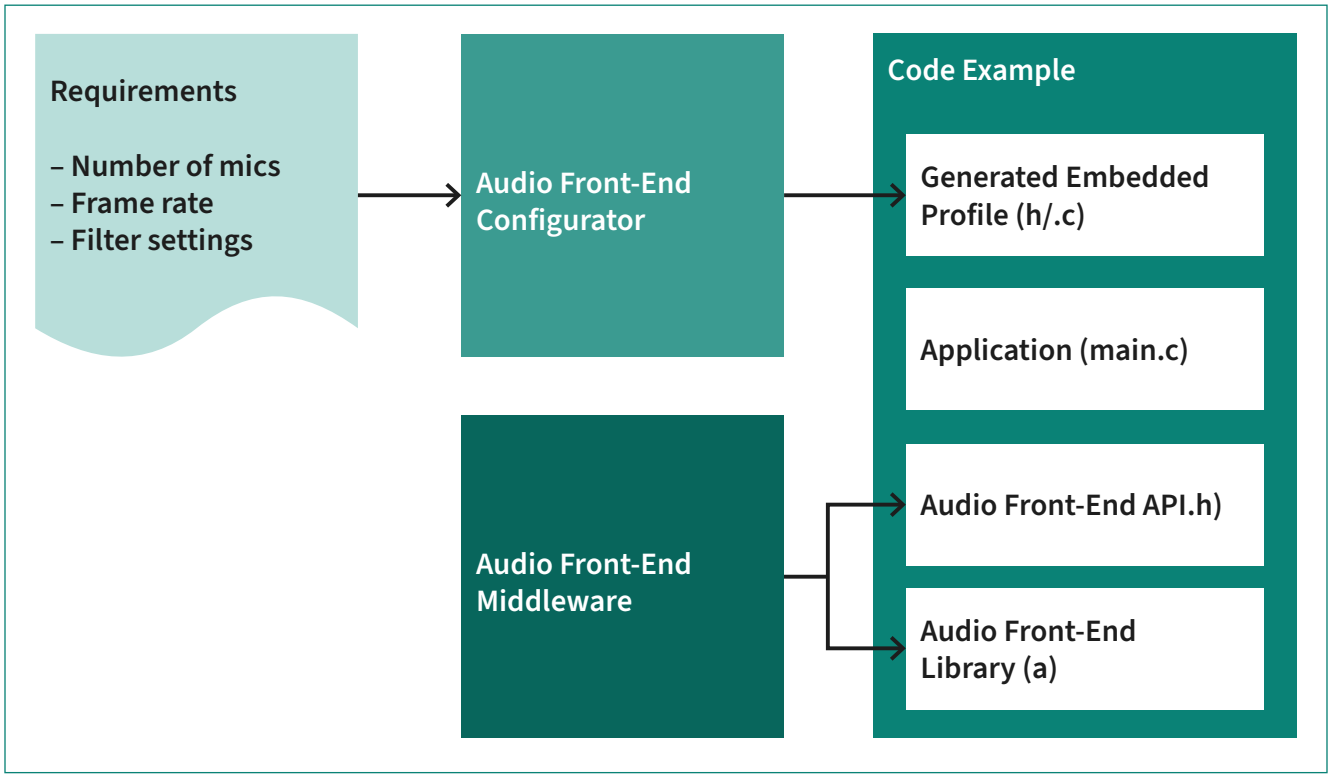


Figure 3 GUI based configurator/tuner

This library simplifies the audio processing for the developer resulting in less time developing the solution as well as better accuracies.

– Staged Voice Control (SVC)

This middleware implements a state machine to simplify voice-based solution for the developer. To do this it uses other libraries such as Low Power Wakeup Word Detection (LPWWD), High Power Wakeup Word Detection (HPWWD), Speech Onset Detection (SOD) etc. The stages can vary based on the type of microphones being used. The following table lists the various stages this middleware can implement, and the blocks used in each stage.

Stages	Description	Using Analog Mics	Using PDM Mics
Stage 0	Listening for acoustic audio	LPPASS: PTC on | ADC off M33: deep-sleep M55: deep-sleep	N/A
Stage 1	Listening for speech	LPPASS: PTC off | ADC on M33: HPF+SOD M55: deep-sleep	Digital mic: powered on M33: SOD M55: deep-sleep
Stage 2	Listening for wake-word (low-power)	LPPASS: PTC off | ADC on M33: HPF+SOD+WWD M55: deep-sleep	Digital mic: powered on M33: SOD+WWD M55: deep-sleep
Stage 3	Listening for wake-word (high-performance) or Reconfirming the wake-word detection	LPPASS: PTC off | ADC on M33: HPF+SOD M55: AFE+WWD	Digital mic: powered on M33: SOD M55: AFE+WWD
Stage 4	Listening for query/command	LPPASS: PTC off | ADC on M33: HPF+SOD M55: AFE+ASR	Digital mic: powered on M33: SOD M55: AFE+ASR
Stage 5	Processing query/command	LPPASS: depends on app M33: processing M55: processing	Digital mic: depends on app M33: processing M55: processing

– Low-Power Wakeup Word Detection (LPWWD)

This library is used to detect a specific word using a pre-trained machine learning model running on NNLite and Cortex®-M33 CPU. This is the first detection of the wakeup word and if detected the audio samples will be passed to the High-Performance Wakeup Word Detection (HPWWD) block for a reconfirmation. The solution is architected in this manner to avoid false positives while keeping the current consumption under check. Using a single HPWWD block will result in frequent wake-up of the high-performance domain of the device resulting in higher power consumption. Whereas, by skipping the reconfirmation using the HPWWD block we risk frequent false positives.

– High-Performance Wakeup Word Detection (HPWWD)

This library is used to detect a specific word using a pre-trained machine learning model running on Cortex®-M55 + Ethos™-U55. It first runs the audio samples through the AFE block to improve its quality. This gives it higher accuracy of detecting the word compared to the LPWWD block but also consumes higher power. If the solution is wall powered, then the LPWWD block can be skipped and only this block used for better latency.

– Automatic Speech Recognition (ASR)

This middleware library is used to recognize certain commands and queries which can then be executed locally on the device. This will run on Cortex®-M55 + Ethos™-U55 in the high-performance domain.

– Audio Streamer

This library implements an internal ring buffer to manage audio data for streaming. This library internally uses the software codec libraries so the developer can use this library to encode and decode the audio data. It can simplify the audio data management for the user.

– Codecs

PSOC™ Edge supports drivers for software and hardware codecs. The software codec library encodes and decodes the audio data for audio compression. It supports various codecs such as LC3, MP3, AAC, OPUS, OGG, SBC, mSBC. The hardware codec library abstracts multiple audio codec chip drivers. It can support codecs from various vendors like Infineon, Cirrus Logic and Asahi Kasei. Native support of these libraries means the developer doesn't have to deal with 3rd party libraries.

– Audio Utils

This is a collection of audio utility functions which can be used by developers as well as other libraries to perform various audio related functions such as:

Speech Onset Detection (SOD):
Speech Onset Detection library detects the beginning of a speech on a series of audio samples.

Sound Meter:
The sound meter library calculates the energy in an audio frame using various algorithm.

Down Sampler:
The down sampler library down samples an audio frame from 48 or 44.1 KHz to 16 KHz.

– Text-to-Speech

The text-to-speech library can synthesize a human voice based on an input text. This library can be used to respond to the queries generated by the user locally thus getting rid of the need to communicate with the cloud.

– Voice Identification

Voice Identification is a biometric technology to identify a person based on his voice. This library can be used with other libraries such as LPWWD or HPWWD to first detect a key word and then to confirm the identity. It can be used for authentication of a user.

– Sound Classification

This library listens to the input audio and the categorizes it based on a target list. For e.g., gun shots, glass break, human voices etc.

2.2.4 Tools

ModusToolbox™ for Machine Learning provides a set of tools and libraries that enables you to rapidly evaluate and deploy Machine Learning models on Infineon MCUs. ModusToolbox™ ML is designed to work seamlessly with the ModusToolbox™ ecosystem and can be added into existing projects to enable inferencing on low-power edge devices. Traditionally, the ML developer needs to segment his models into a DSP workload, neural net workload and CPU workload. This is then cross-compiled and hand-optimized to run on the target embedded platform which can be time consuming.

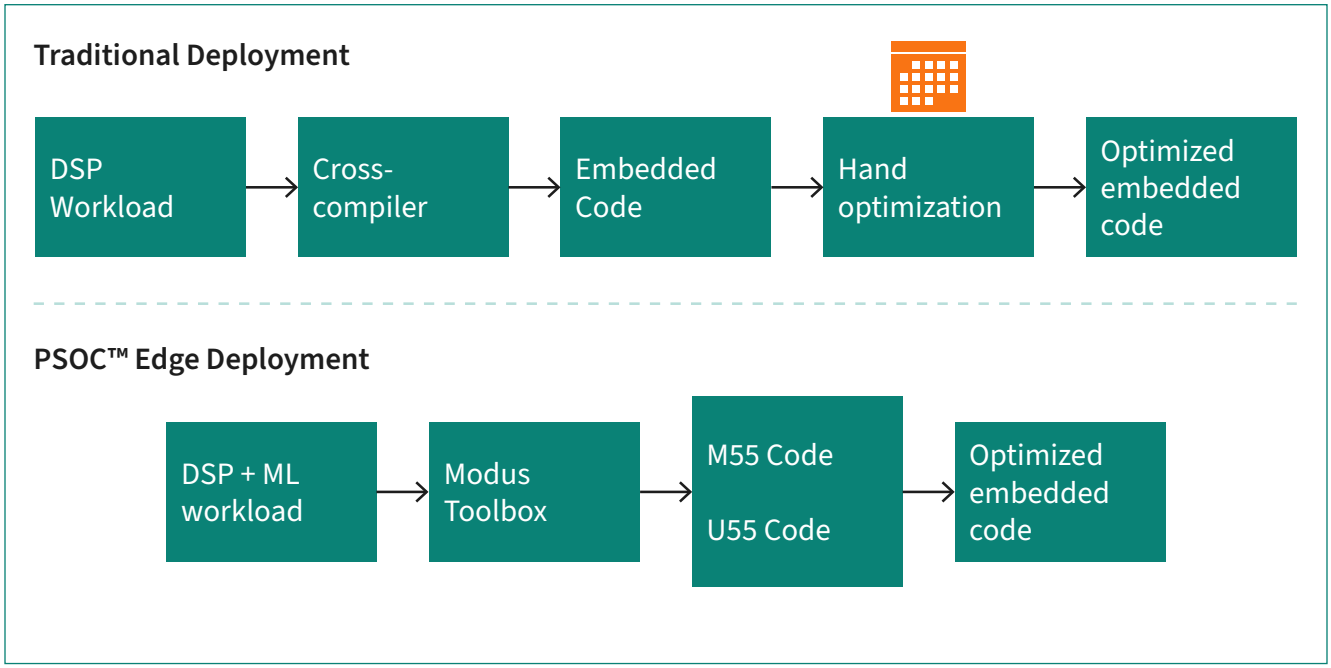


Figure 4 ModusToolbox™ for Machine Learning Deployment

ModusToolbox™ for Machine Learning simplifies this process by providing tools to divide the workload between the Cortex®-M55 or Ethos™-U55 cores thus eliminating any hand optimization and reducing development time. These tools support .tflite and .H5 model formats, Input data quantization level of 32-bit float and 16/8bit integer and also give cycle and memory estimation. The libraries also support CMSIS-NN which contains highly optimized and performant kernels that accelerate a subset of operators in the TFLM framework.

3 Audio and Voice Use Cases

3.1 Battery Powered Local Voice

In this use case we will typically have an always-on voice assistant on the device with limited capabilities. The user will first speak a wakeup word followed by a command or a query. The LPPASS first detects the audio using the comparator. If audio is detected it turns on the ADC to collect audio samples. Once enough samples are available it passes it on to the Cortex®-M33 core which filters the data and then runs a machine learning algorithm on the NNLite block to detect the pre-determined wakeup word. If it detects the wakeup word it passes it to the Cortex®-M55 core which first processes the audio samples and then runs a more sophisticated machine learning algorithm on the Ethos™-U55 block to confirm the wakeup word. This process is depicted in the following diagram.

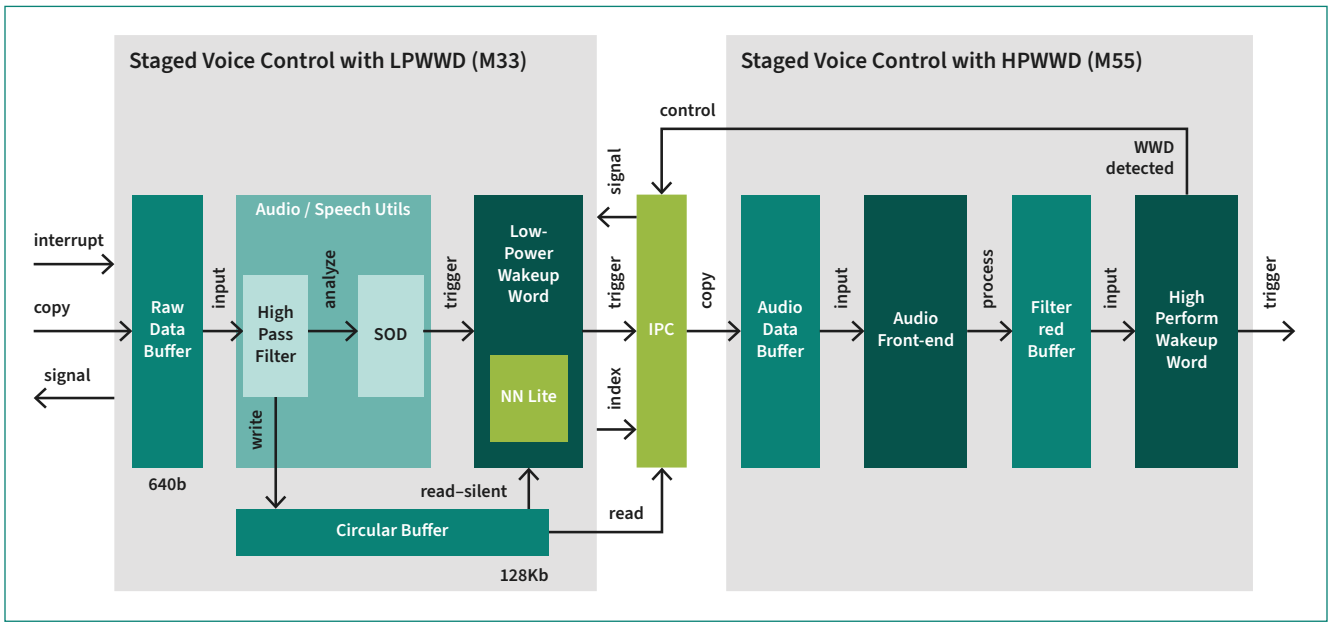


Figure 5 Battery Powered Local Voice Use Case

Once the wakeup word is detected, the command that follows it is then processed by the Cortex®-M55 and Ethos™ U55. The Automatic Speech Recognition block helps with this. The wakeup word and the command are processed in various stages as shown below. The states of the peripherals in various stages are also depicted below.

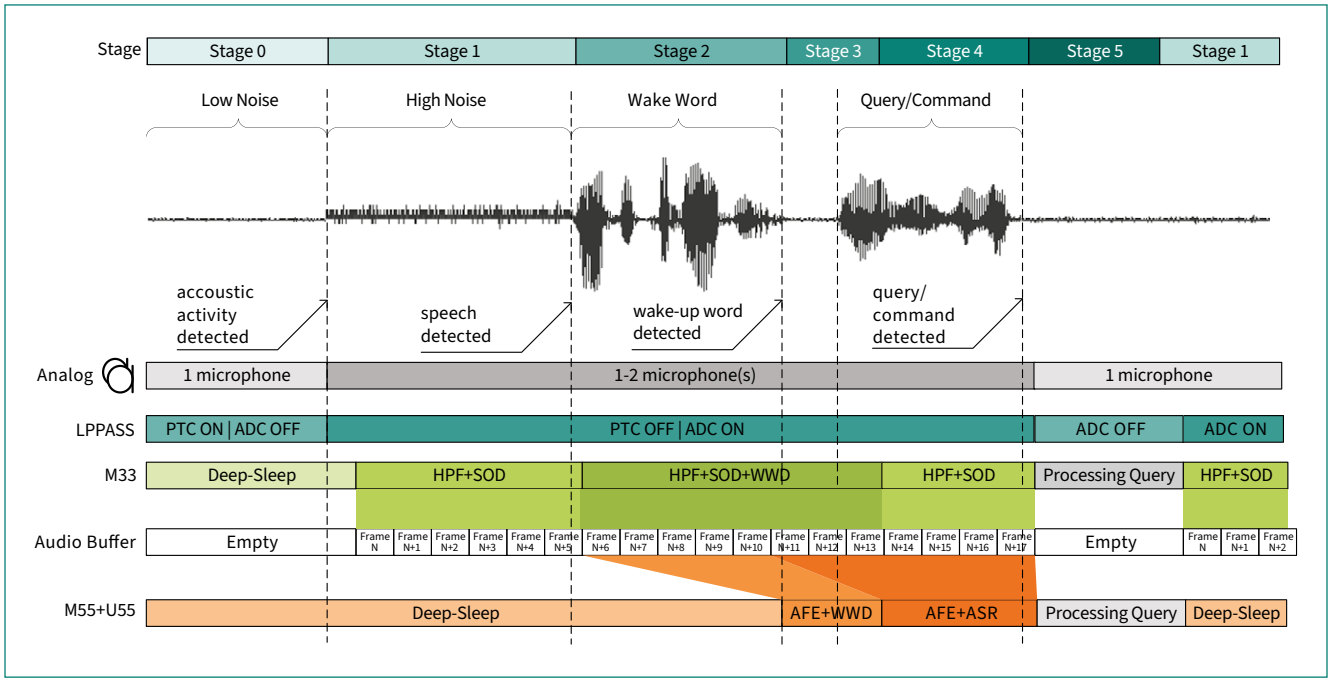


Figure 6 Battery Powered Local Voice Use Case

Once we have processed the command or the query, we need to respond to it. If it is a command like play or pause music, the application can start and stop the audio output. If it is a query then we search for the answer and then use the Text-to-Speech block to synthesize a human voice to play the answer. The output path is as depicted below.

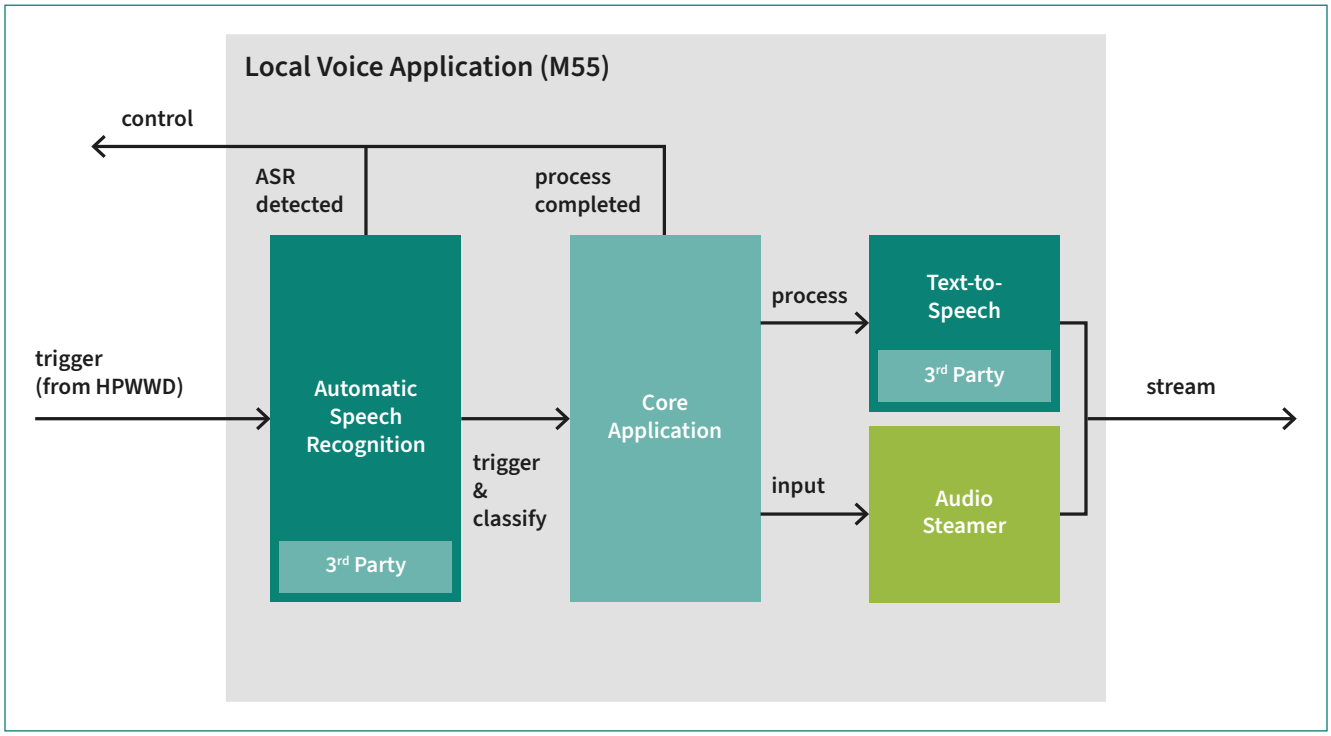


Figure 7 Battery Powered Local Voice Use Case

If the audio is being played over a loudspeaker then this data is also fed into the audio echo cancellation block. If the audio is sent over to Bluetooth earbuds, then AEC is not required.

3.2 Battery Powered Cloud Voice

For cloud voice, the process to detect the wakeup word remains the same as for local voice. But instead of the command being processed locally it is sent to the cloud for processing. An example for this is the Amazon Mobile Accessory SDK or Alexa Voice Services. The following image shows this scenario.

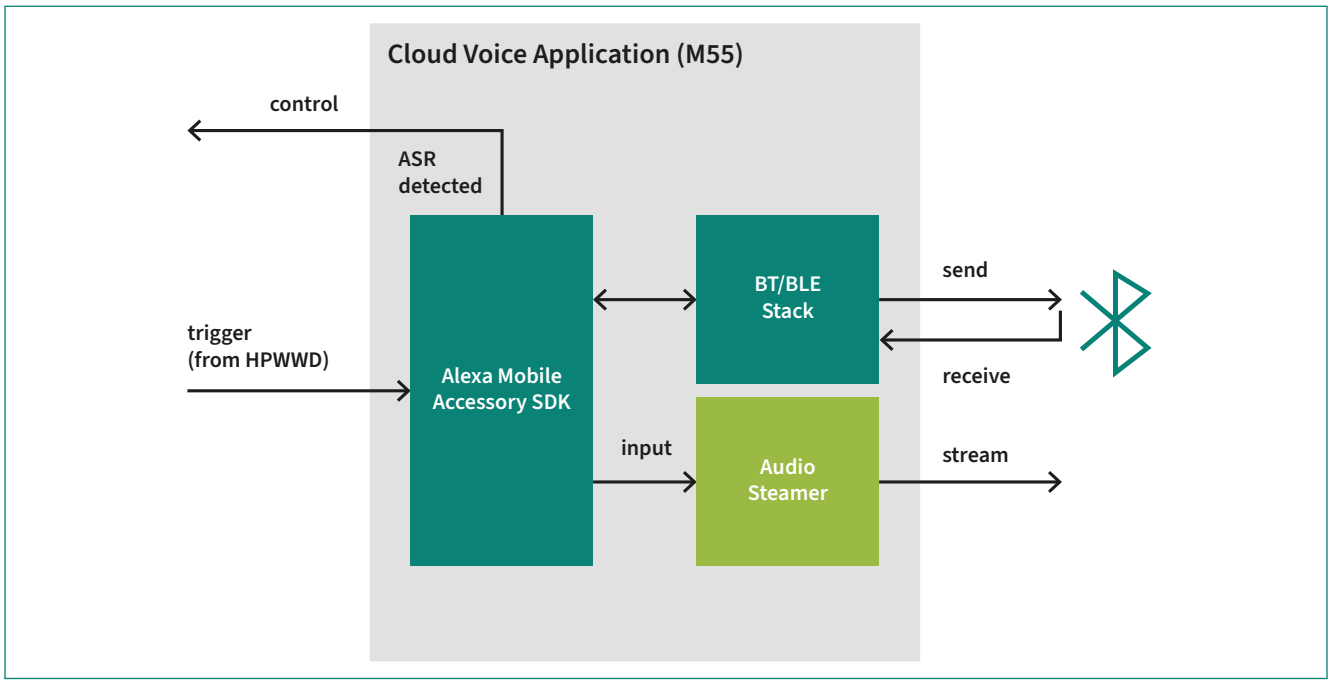


Figure 8 Abattery Powered Cloud Voice Use Case

3.3 Mains Powered Voice

Most of the system architecture for mains powered use case remains the same as the battery powered use case. Since we are not concerned about the battery life we can skip the low power wakeup word detect and pass the audio data to the high performance wakeup word detection block. This will reduce the latency of the system.

3.4 Voice Identification

In this use case we want to identify the person speaking a wakeup word. This is done using the LPWWD and HPWWD blocks along with a voice identification algorithm running on the Cortex®-M55 and Ethos™ U55. This can be used as a security feature to authenticate a user.

4 Summary

The hardware blocks within PSOC™ Edge augmented by advanced software algorithms can enable implementation of advance audio and voice solution on a small cost-effective MCU platform which are typically reserved for much higher power, cost, and complexity devices. It also reduces the overall system power consumption thus enabling always-on voice on new form factor devices. The ability to process voice locally and respond using local text-to-speech library means it can work without an active internet connection.

References

[1] https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-u55

[2] https://developer.arm.com/AI%20and%20ML#aq=%40navigationhierarchiescategories%3D%3D%22AI%2FML%22&numberOfResults=48

[3] https://www.arm.com/-/media/Files/pdf/ML%20on%20Arm/Arm_Ethos_U55_Product_Brief.pdf?revision=921d33c8- d166-4fb6-abf8-92ec306a0eeb&rev=921d33c8d1664fb6abf892ec306a0eeb&hash=1D3B17357D-02451F7B4788F 5BAF00F2F

[4] https://www.arm.com/-/media/Files/pdf/ethos/Arm_Accelerating_ML_Compute_for_Embedded_Market_white_paper1.pdf?revision=8772b5c9-89fa-420b-92a3-0d77e91c4597&rev=b5af90cebeb748b3ba54e5e71a988c7b&hash=673095A6389EE3DA9E7C1E05FA6C83A2

[5] https://www.infineon.com/cms/en/design-support/tools/sdk/modustoolbox-software/?gclid=CjwKCAjw-IWkBhBTEiwA2exyOzBwjtIFYIzxSJQrsXGqVRpzWACcslwmCTcZOMU9WhGV0T2IWgqpoxoCIUAQAvD_BwE&gclsrc=aw.ds

[6] https://www.infineon.com/cms/en/design-support/tools/sdk/modustoolbox-software/modustoolbox-machine-learning/

Published by
Infineon Technologies AG
Am Campeon 1-15, 85579 Neubiberg
Germany

Public

Date: 6/2024

Stay connected!

Scan QR code and explore offering
www.infineon.com

Please note!
This Document is for information purposes only and any information given herein shall in no event be regarded as a warranty, guarantee or description of any functionality, conditions and/or quality of our products or any suitability for a particular purpose. With regard to the technical specifications of our products, we kindly ask you to refer to the relevant product data sheets provided by us. Our customers and their technical departments are required to evaluate the suitability of our products for the intended application.

We reserve the right to change this document and/or the information given herein at any time.

Additional information
For further information on technologies, our products, the application of our products, delivery terms and conditions and/or prices, please contact your nearest Infineon Technologies office (www.infineon.com).

Warnings
Due to technical requirements, our products may contain dangerous substances. For information on the types in question, please contact your nearest Infineon Technologies office.

Except as otherwise explicitly approved by us in a written document signed by authorized representatives of Infineon Technologies, our products may not be used in any life-endangering applications, including but not limited to medical, nuclear, military, life-critical or any other applications where a failure of the product or any consequences of the use thereof can result in personal injury.