European Commission

Horizon 2020
European Union funding
for Research & Innovation

Big Data technologies and extreme-scale analytics

MARVEL

Multimodal Extreme Scale Data Analytics for Smart Cities Environments

D4.1: Optimal audio-visual capturing, analysis and voice anonymisation – initial version[†]

Abstract: This report focuses on the initial version of the optimal audio-visual capturing, analysis and voice anonymisation. In the field of audio capturing, this document gives detailed information about the hardware developed and used in the scope of this project. Some consist of only the microphone and a simple connector, while others come with a processing unit that allows audio capturing via USB, streaming to a cloud via WiFi and even Edge AI processing. First versions of different data acquisition installations using the already available audio devices in different scenarios are presented, along with the first results of those experiments. This section concludes with information about plans for pre-processing for data analysis through the usage of Edge AI techniques. The second section of the document is about devAIce SDK, a modular technology optimised to work on cross-platforms and contains several AI models such as the Voice Activity Detection (VAD) tool as well as the feature extraction toolkit, openSMILE. It is used for audio analytics and sound event detection, acoustic scene classification and speech analysis which is a fundamental step to ensure privacy compliance and is designed to be implemented on high-end edge devices. Furthermore, an Android app used to record environmental acoustics and user annotations, SensMiner, is introduced.

[†] *The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337.*

Contractual Date of Delivery	31/12/2021
Actual Date of Delivery	30/12/2021
Deliverable Security Class	Public
Editor	*Elfi Fertl (IFAG)*
Contributors	IFAG, AU, AUD, CNR, UNS, ZELUS, FBK, TAU, GRN, FORTH
Quality Assurance	*Bahaeddine Abrougui (AUD)* *Dora Kallipolitou (ZELUS)*

The *MARVEL* Consortium

Part. No.	Participant organisation name	Participant Short Name	Role	Country
1	FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS	FORTH	Coordinator	EL
2	INFINEON TECHNOLOGIES AG	IFAG	Principal Contractor	DE
3	AARHUS UNIVERSITET	AU	Principal Contractor	DK
4	ATOS SPAIN SA	ATOS	Principal Contractor	ES
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR	Principal Contractor	IT
6	INTRASOFT INTERNATIONAL S.A.	INTRA	Principal Contractor	LU
7	FONDAZIONE BRUNO KESSLER	FBK	Principal Contractor	IT
8	AUDEERING GMBH	AUD	Principal Contractor	DE
9	TAMPERE UNIVERSITY	TAU	Principal Contractor	FI
10	PRIVANOVA SAS	PN	Principal Contractor	FR
11	SPHYNX TECHNOLOGY SOLUTIONS AG	STS	Principal Contractor	CH
12	COMUNE DI TRENTO	MT	Principal Contractor	IT
13	UNIVERZITET U NOVOM SADU FAKULTET TEHNICKIH NAUKA	UNS	Principal Contractor	RS
14	INFORMATION TECHNOLOGY FOR MARKET LEADERSHIP	ITML	Principal Contractor	EL
15	GREENROADS LIMITED	GRN	Principal Contractor	MT
16	ZELUS IKE	ZELUS	Principal Contractor	EL
17	INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK	PSNC	Principal Contractor	PL

Document Revisions & Quality Assurance

Internal Reviewers

1. *Bahaeddine Abrougui, AUD*
2. *Dora Kallipolitou, ZELUS*

Revisions

Version	Date	By	Overview
1.1.0	29/12/2021	Elfi Fertl, PC	Final review and approval
1.0.0	22/12/2021	Elfi Fertl	2nd Reviewer (ZELUS) comments addressed
0.9.0	16/12/2021	Antonio Escobar	1st Reviewer (AUD) comments addressed
0.8.0	10/12/2021	Antonio Escobar	First version for review
0.7.0	06/12/2021	Elfi Fertl	All contributions compiled
0.0.4	19/11/2021	Elfi Fertl	First draft without input from partners.
0.0.3	16/11/2021	Antonio Escobar	Updated ToC.
0.0.2	20/10/2021	Elfi Fertl	Updated ToC.
0.0.1	04/10/2021	Antonio Escobar	ToC.

Disclaimer

The work described in this document has been conducted within the MARVEL project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337. This document does not reflect the opinion of the European Union, and the European Union is not responsible for any use that might be made of the information contained therein.

This document contains information that is proprietary to the MARVEL Consortium partners. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to any third party, in whole or in parts, except with prior written consent of the MARVEL Consortium.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
ARM	Advanced RISC Machines
CPU	Central Processing Unit
DL	Deep Learning
DMT	Decision-Making Toolkit
E2F2C	Edge-to-Fog-to-Cloud
EC	European Commission
GA	Grand Agreement
GND	System Ground
GPS	Global Positioning System
GPU	Graphics Processing Unit
I2S	Inter-IC Sound
IC	Integrated Circuit
IDE	Integrated Development Environment
IoT	Internet of Things
JSON	JavaScript Object Notation
JTAG	Joint Test Action Group (Debug Interface)
KPI	Key Performance Indicator
LED	Light Emitting Diode
LSTM	Long Short-Term Memory
MCU	Microcontroller Unit
MEMS	Micro-Electromechanical System
ML	Machine Learning
OS	Operating System
PCB	Printed Circuit Board
PCM	Pulse Code Modulation
PDM	Pulse-density Modulation
SDK	Software Development Kit
SNR	Signal-to-Noise Ratio
SWD	Single Wire Debug
UAR	Unweighted Average Recall
UI	User Interface
USB	Universal Serial Bus

VAD	Voice Activity Detection
VDD	Drain Power Voltage
WAV	Waveform Audio File
WiFi	Wireless Fidelity
WP	Work Package
ZIF	Zero Insertion Force

Executive Summary

This deliverable is the initial version of the optimal audio-visual capturing, analysis and voice anonymisation. The document reports the partial progress of Task 4.1 and Task 4.2 achieved in M12, towards the D4.4 optimal audio-visual capturing, analysis and voice anonymisation. Both Tasks run until M24, which is also the due date of D4.4.

The goals of T4.1 and T4.2 contribute to Objective 1 of the project, i.e., *"Leverage innovative technologies for data acquisition, management and distribution to develop a privacy-aware engineering solution for revealing valuable and hidden societal knowledge in a smart city environment"* and are the following:

- Describe innovations performed by analogue and digital microphones that are based on MEMS technology.
- Utilise the devAIce platform for audio analysis and feature extraction.

In the first part of the document, the microphone used for audio data acquisition, the XENSIV^{TM} IM69D130 MEMS microphone, is described in detail. Information about the provided features like operating parameters, frequency response, performance and acoustic characteristics is given. Several boards to connect the microphone were provided and several audio data pre-processing boards were introduced. The advantage of the boards that include audio pre-processing functionalities is that they can be easily connected to a processing unit via USB and send the data using the serial protocol I2S (Inter-IC Sound). The Audiohub – Nano 8 will come with a WiFi connector that allows sending the audio data directly to a cloud and a strong microcontroller that can run Machine Learning (ML) applications. Findings and outcomes of the first data acquisition experiments in different locations, scenarios, and setups of the devices are described. Finally, this section includes justifications for pre-processing and data analysis through the usage of Edge AI techniques.

The second part of the document is dedicated to audio data analysis and voice anonymisation. For sound event detection, acoustic scene classification, speech analysis, and audio feature extraction the devAIce platform was introduced. It is an SDK (Software Development Kit) that runs on high-end computers as well as medium edge devices and comes with interfaces in Python, iOS, Android and C. It includes the feature extraction toolkit openSMILE that provides a stream of features from an audio stream, ready to be fed into an ML model. This will provide input for SmartViz and ultimately the Decision-Making Toolkit (DMT). DevAIce features a Voice Activity Detection (VAD) module, which has been improved in the scope of MARVEL. This document provides information about the improvements achieved. The integration of the new VAD into devAIce is currently ongoing. In addition, a music detection tool was integrated in devAIce that allows music and speech detection simultaneously. Moreover, SensMiner was updated, an Android app used to record environmental acoustics and user annotations.

D4.1 documents audio data acquisition, pre-processing and processing hardware and software, which provide the beginning of an optimal audio data acquisition, pre-processing and processing pipeline. Acquired audio data is directly processed such that just the necessary information in form of features is passed on; a stream of condensed information. In future development, when the implementation of ML models on edge devices is finished only the output of the ML models will be given to a central processing unit. Thereby reducing the amount of data that is being transmitted, complying with privacy constraints, and reducing the overall power consumption.

1 Introduction

1.1 Purpose and scope

This deliverable reports the partial progress, of T4.1 (Optimised audio capturing through MEMS devices, M6 – M24) up to M12 and T4.2. (devAIce platform for audio-visual analysis and voice anonymisation, M6 – M24). Both tasks will be completed and their final update will be delivered, in the D4.4 (Optimal audio-visual capturing analysis and voice anonymisation), due in M24.

In T4.1, the MEMS microphone intended to be used throughout the project is introduced. The first platforms to evaluate its capabilities are already available. A novel platform with advanced features (up to eight microphones and Edge AI capabilities) is under development. First evaluation results with the USB Audio platforms using the MEMS devices are available. In T4.2, the first results with the VAD toolkits within the context of the devAIce platform as well as the SensMiner app, are provided. Both tasks will report their complete results in D4.4.

1.2 Relation to other work packages, deliverables and activities

The platforms introduced and developed in T4.1 will be used to implement and deploy the models defined in T3.4 (Adaptive E2F2C distribution and optimisation of AI tasks). Close cooperation between the partners from those tasks is required to evaluate the required computing capabilities for the deployment of the ML networks directly at the edge. Edge platforms featuring the MEMS microphones with different capabilities are provided to the partners of WP3, enabling the deployment of both relatively complex networks in powerful embedded computing platforms featuring Linux-based OS, like Raspberry Pi, and simple models in low-power microcontroller platforms, like a PSoC6.

The devAIce platform and associated toolkits, developed in T4.2, will fulfil the requirements defined for the project experiments in T1.3 (Experimental protocol – real life societal trial cases in smart cities environments). Furthermore, the synergies between the platform and the techniques defined in T3.3 (Multimodal audio-visual intelligence) and T3.1 (AI-based methods for audio-visual data privacy) are exploited.

The platforms will be used by the pilots for data acquisition. In this sense, both tasks are additionally linked to WP6.

1.3 Contribution to WP4 and project objectives

T4.1 fulfils one of the basic objectives of WP4 by describing the innovations performed by analogue and digital microphones based on MEMS technology. Furthermore, it contributes to the global project Objective 1 (Leverage innovative technologies for data acquisition, management and distribution to develop a privacy-aware engineering solution for revealing valuable and hidden societal knowledge in a smart city environment), particularly in KPI-01-E1-2 (Increase of data throughput and decrease access latency by 10%). This is done by developing resourceful and autonomous edge platforms, able to perform the processing directly in the sensing device, reducing latency and bandwidth usage.

T4.2 also contributes to project Objective 1, both in KPI-O1-E3-1 (Number of incorporated safety mechanisms (e.g., for privacy, voice anonymisation) ≥ 3) by means of the features implemented in the devAIce platform.

2 Audio capturing with MEMS devices

This section is mainly related to T4.1 (Optimised audio capturing through MEMS devices). The first subsection (2.1), describes the features and provides a comprehensive description of the MEMS microphones provided by IFAG. Further information is given about operating parameters, physical description, frequency response, performance, and other relevant features of the microphones themselves as well as the connecting boards.

In the recent past, microphones have become increasingly important as the number of audio features and applications are rising, not only in mobile phones; but also, with IoT devices like smart speakers, that require high-quality audio capturing to effectively perform AI processing. With the XENSIV^{TM} MEMS microphones, IFAG offers a wide variety of microphones from low cost to high performance.

Since the 60th more and more sensors are being replaced by MEMS sensors. Advantages of MEMS sensors compared to traditional microphones are low bias and actuation voltage operation, their small form factor and their high SNR.

IFAG's MEMS microphones provide the audio signal in PDM format. PDM is a system for representing a sampled signal as a stream of single bits and is used to represent mono or stereo audio in digital format. Its main characteristics are a high sampling frequency and that it uses single bits.

The microphones can either be directly soldered to a PCB or connected via a processing IC, that converts the raw PDM signal to I2S, which can then be easily processed with a typical microcontroller, like an Arduino. I2S is a standard serial bus interface used to connect audio devices. It consists of the three lines: continuous serial clock, word select (left or right channel) and serial data. They are used to synchronise the timing of the communicating partners and transmit the audio data. Up to two I2S signals can be multiplexed in the same data line.

In the second subsection (2.2), different processing boards are presented, enabling the audio data gathering in different higher-level formats; further converting the I2S signal to USB Audio or to a WAV stream sent via WiFi. A board that streams mono or stereo audio data via USB is already available. A further board that can simultaneously stream audio data from 4 channels via USB has just been developed, and currently, a board with 8 microphones is in the development process. The board with 8 microphones comes with a rather powerful microcontroller that can be programmed to implement AI processing directly at the edge.

2.1 Investigation of operating parameters of MEMS microphones

The XENSIV^{TM} IM69D130 MEMS microphone is the basis of the audio data acquisition. All evaluation Kits feature the latest state-of-the-art IM69D130 MEMS microphone, as depicted in Figure 1. It is the newest MEMS microphone produced by IFAG for consumer electronics. Figure 2 shows its block diagram.

More specific information about operating parameters, frequency response, performance and other acoustic features can be found in Table 1. Furthermore, a comparison of the key features of the current microphone with respect to previous IFAG's model is presented in Table 2, displaying the progress with respect to the state of the art (IM69D120).

There are analog microphones available with lower power consumption and higher sensitivity (IM73A135), but the integrated low-power analog-to-digital conversion is more suitable for the envisioned IoT applications, in terms of size cost, power, simplicity and flexibility. Figure 3 shows the mechanical structure of the microphone. The acoustic hole of the microphone is located on its bottom side. The digital microphones provide the signal in PDM format.

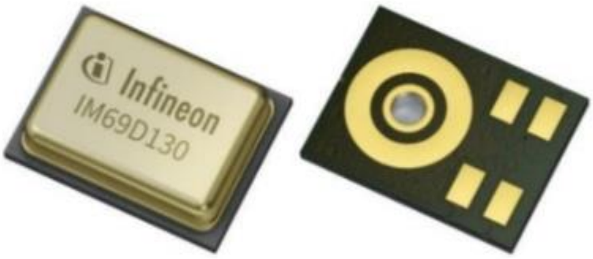


Figure 1. Latest IFAG MEMS microphone IM69D130

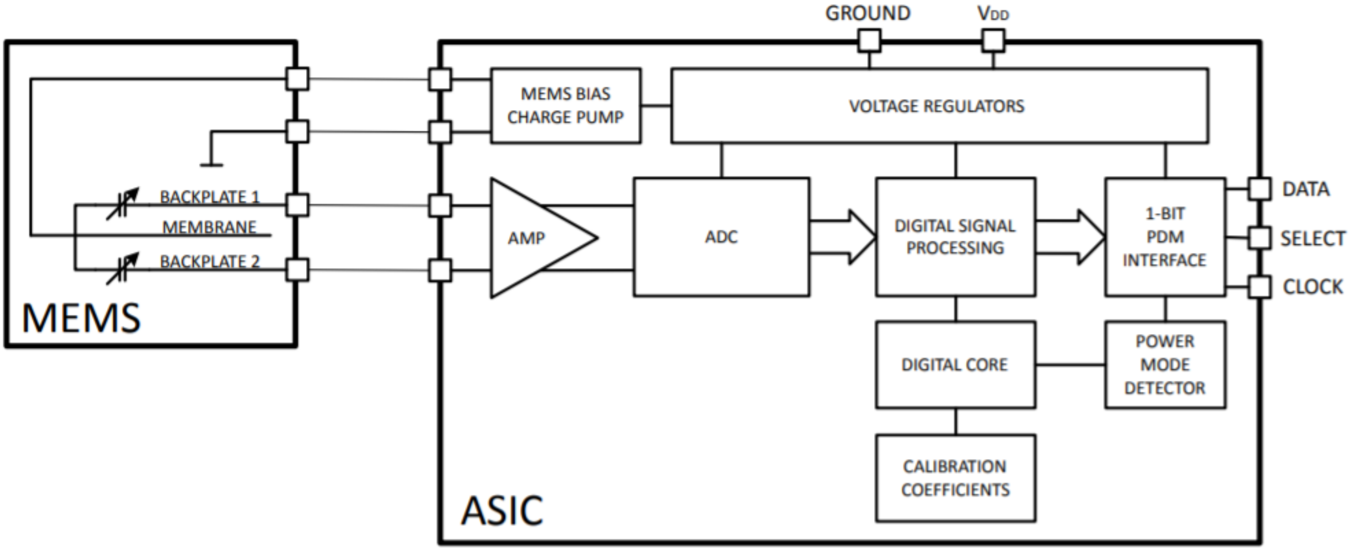


Figure 2. IM69D130 block diagram

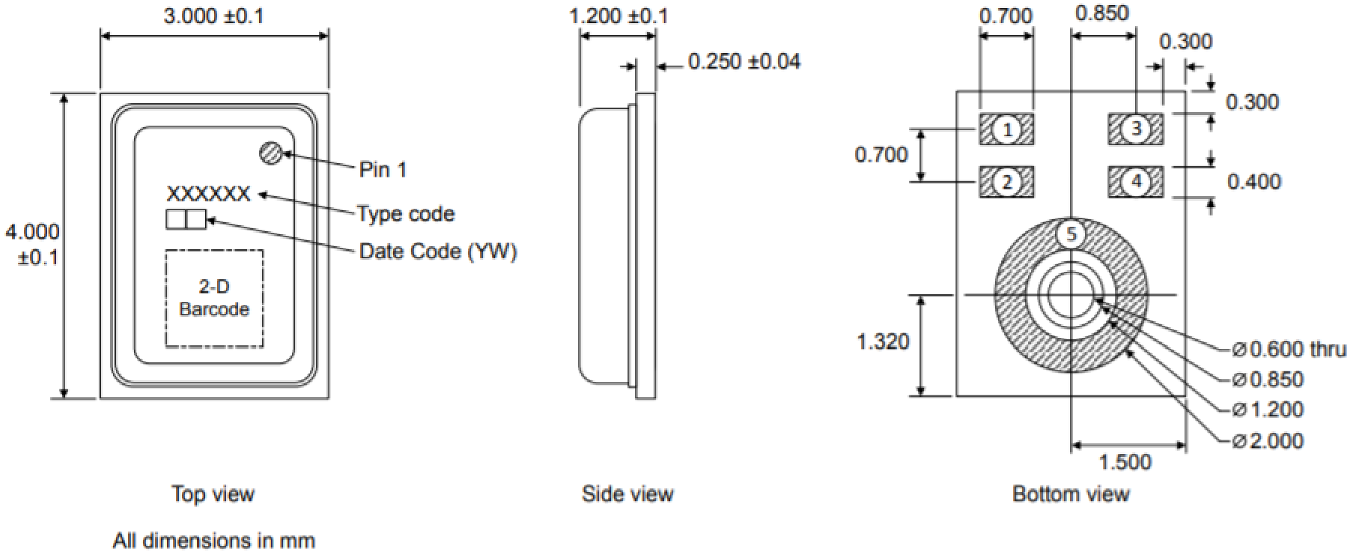
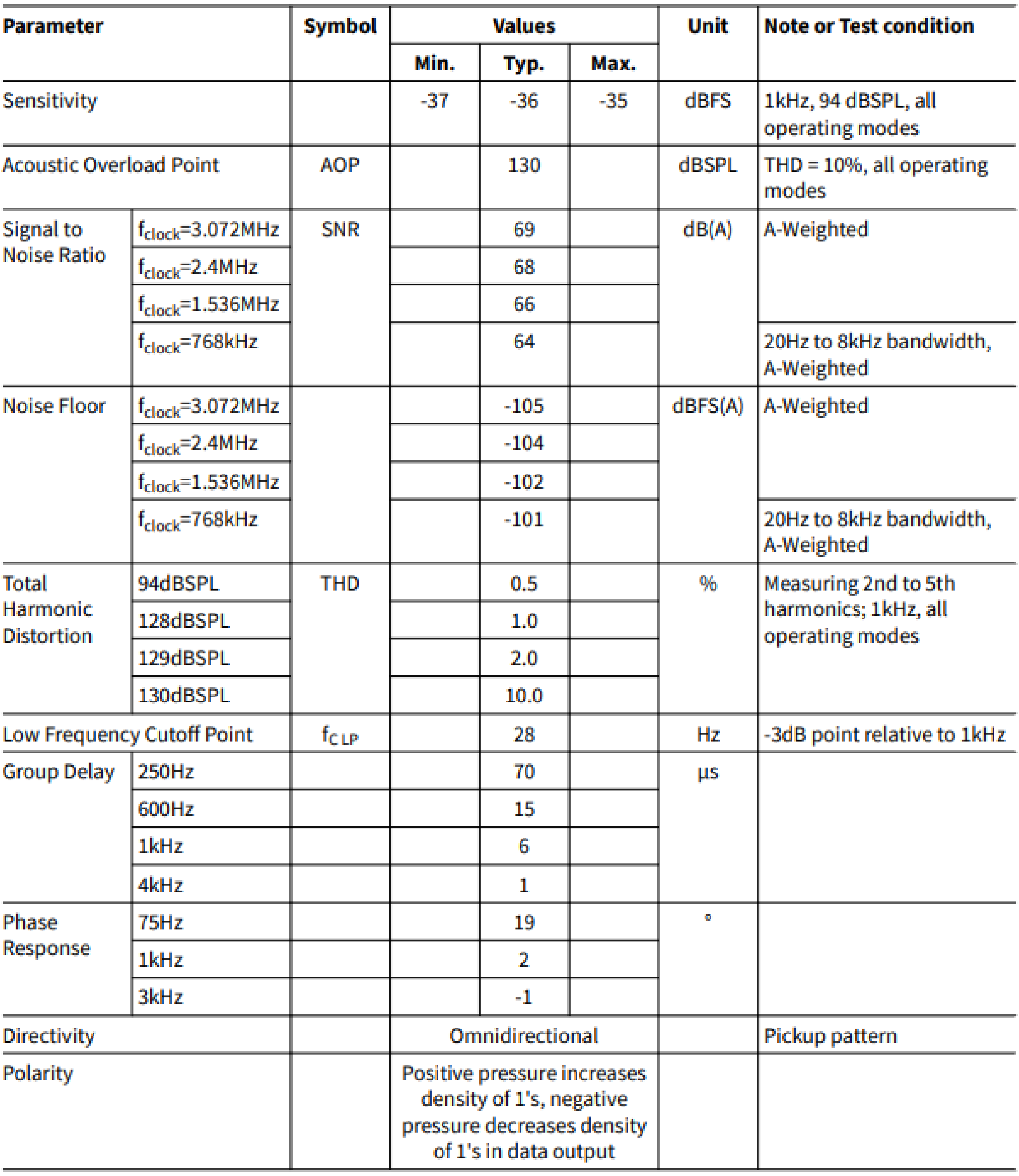


Figure 3. Package dimensions of IM69D130

Table 1: Acoustic characteristics IM69D130



Parameter		Symbol	Values			Unit	Note or Test condition
			Min.	Typ.	Max.		
Sensitivity			-37	-36	-35	dBFS	1kHz, 94 dBSPL, all operating modes
Acoustic Overload Point		AOP		130		dBSPL	THD = 10%, all operating modes
Signal to Noise Ratio	f_{clock}=3.072MHz	SNR		69		dB(A)	A-Weighted
	f_{clock}=2.4MHz			68			
	f_{clock}=1.536MHz			66			
	f_{clock}=768kHz			64			20Hz to 8kHz bandwidth, A-Weighted
Noise Floor	f_{clock}=3.072MHz			-105		dBFS(A)	A-Weighted
	f_{clock}=2.4MHz			-104			
	f_{clock}=1.536MHz			-102			
	f_{clock}=768kHz			-101			20Hz to 8kHz bandwidth, A-Weighted
Total Harmonic Distortion	94dBSPL	THD		0.5		%	Measuring 2nd to 5th harmonics; 1kHz, all operating modes
	128dBSPL			1.0			
	129dBSPL			2.0			
	130dBSPL			10.0			
Low Frequency Cutoff Point		$f_{C\ LP}$		28		Hz	-3dB point relative to 1kHz
Group Delay	250Hz			70		µs	
	600Hz			15			
	1kHz			6			
	4kHz			1			
Phase Response	75Hz			19		°	
	1kHz			2			
	3kHz			-1			
Directivity			Omnidirectional				Pickup pattern
Polarity			Positive pressure increases density of 1's, negative pressure decreases density of 1's in data output				

Table 2: Comparison of key features of IM69D130 to other IFAG MEMS microphones

Product	OPN	Package	Current consumption	Sensitivity	Signal to noise [dB]	Supply voltage [V]
IM69D130	IM69D130V01XTSA1	LLGA-5-1	980 µA	-36 dBFS	69	1.62-3.6
IM69D120	IM69D120V01XTSA1	LLGA-5-1	980 µA	-26 dBFS	69	1.62-3.6
IM73A135	IM73A135V01XTSA1	LLGA-5-2	170 µA @ 2.75 V, 70µA @ 1.6V	-38 dBV	73	1.52-3.0

For quick evaluation purposes, the microphones can be seamlessly connected to an edge device, without directly soldering them, by using either the evaluation kit shown in Figure 4 or the microphone shield presented in Figure 5. Both are described in the following subsections.

2.1.1 EVAL_IM69D130_FLEXKIT

The EVAL_IM69D130_FLEXKIT includes 5 flex boards with five microphones, ready to be evaluated via the flex connector (6-position ZIF). Additionally, it includes an adapter board in case a more classical pin-based interface is desired. It provides the raw PDM signal coming for the digital MEMS microphone, so normally it is connected to a I2S to PDM or I2S to PCM converter to reduce the complexity of the interface.

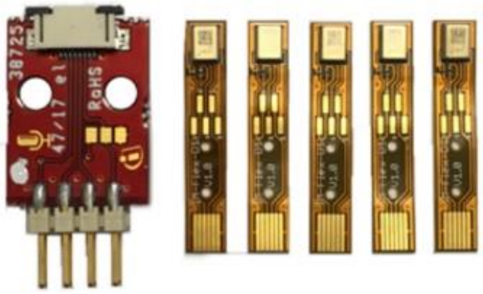


Figure 4. EVAL_IM69D130_FLEXKIT

Table 3: EVAL_IM69D130_FLEXKIT pin configuration

Pin number	Symbol	Digital	Analog
1	V	VDD	VDD
2	D	Data	OUT +
3	C	Clock	OUT -
4	S	Select	Not used
	Back side	GND	GND

2.1.2 IM69D130 Microphone Shield2Go

The IM69D130 Microphone Shield2Go (Figure 5 and Figure 6) features 2 IM69D130 microphones with a supply voltage of 3.3V, including an on-board dual-channel PDM to I2S converter to simplify the interface with a standard microcontroller. Using the IFAG My IoT Adapters, it can be connected to either a Raspberry Pi (Figure 7) or an Arduino-based microcontroller (Figure 8). The included software is fully integrated into the Arduino IDE.

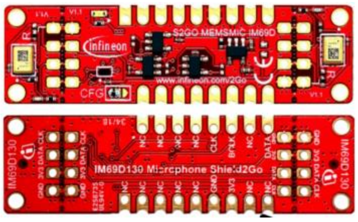


Figure 5. IM69D130 Microphone Shield2Go

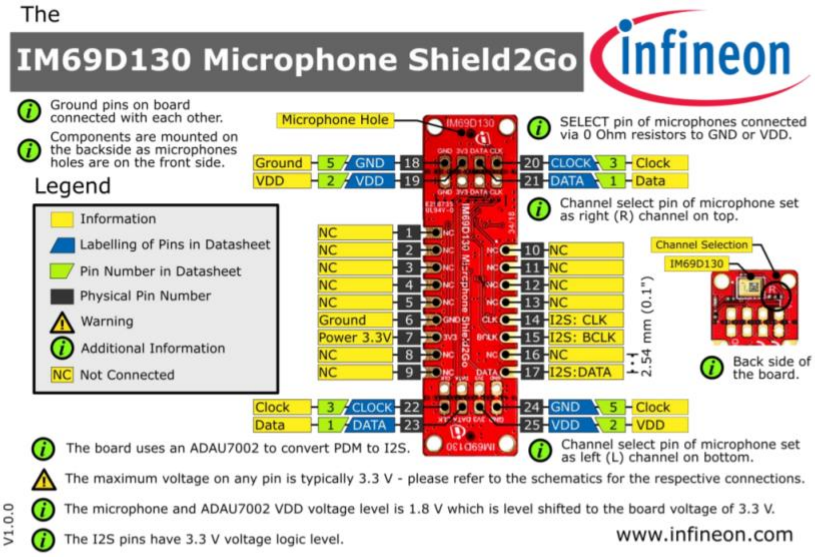

The
IM69D130 Microphone Shield2Go
Infineon
Ground pins on board connected with each other.
Components are mounted on the backside as microphones holes are on the front side.
Legend
Information
Labelling of Pins in Datasheet
Pin Number in Datasheet
Physical Pin Number
Warning
Additional Information
NC Not Connected
Microphone Hole
SELECT pin of microphones connected via 0 Ohm resistors to GND or VDD.
Ground 5 GND 18
VDD 2 VDD 19
20 CLOCK 3 Clock
21 DATA 1 Data
Channel select pin of microphone set as right (R) channel on top.
NC 1
NC 2
NC 3
NC 4
NC 5
Ground 6
Power 3.3V 7
NC 8
NC 9
10 NC
11 NC
12 NC
13 NC
14 I2S: CLK
15 I2S: BCLK
16 NC
17 I2S:DATA
2.54 mm (0.1")
Channel Selection
IM69D130
Back side of the board.
Clock 3 CLOCK 22
Data 1 DATA 23
24 GND 5 Clock
25 VDD 2 VDD
The board uses an ADAU7002 to convert PDM to I2S.
Channel select pin of microphone set as left (L) channel on bottom.
The maximum voltage on any pin is typically 3.3 V - please refer to the schematics for the respective connections.
The microphone and ADAU7002 VDD voltage level is 1.8 V which is level shifted to the board voltage of 3.3 V.
The I2S pins have 3.3 V voltage logic level.
V1.0.0
www.infineon.com

Figure 6. IM69D130 Microphone Shield2Go pin configuration

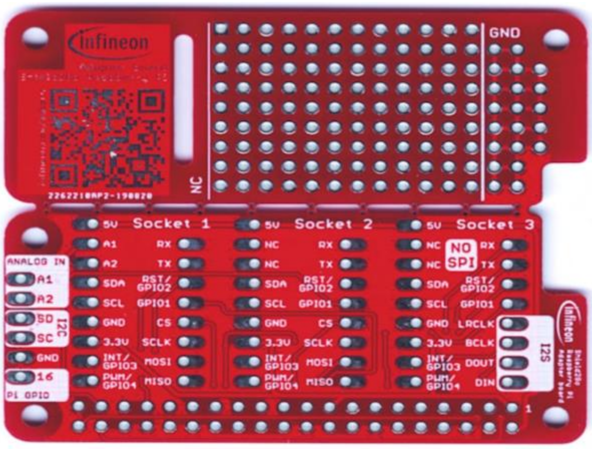


Figure 7. IFAG My IoT adapter for Raspberry Pi

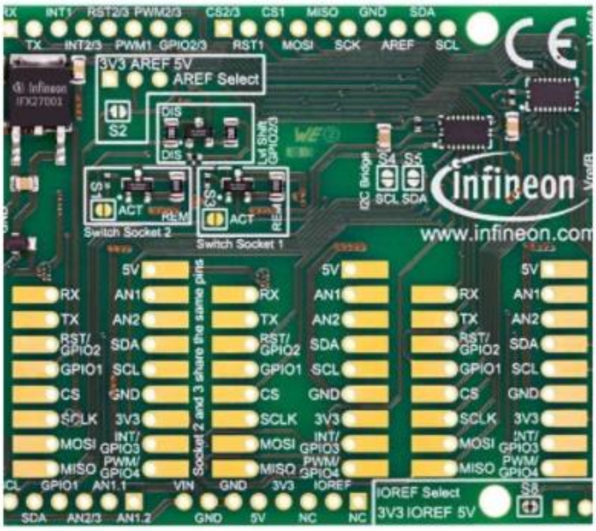


Figure 8. IFAG My IoT adapter for Arduino

2.2 Edge devices featuring MEMS microphones

With the interface boards introduced in the previous sections, the microphones can be evaluated both with standard Arduino- and Raspberry Pi-based platforms. Additionally, several custom boards are being designed, enabling advanced functionality, like the deployment of multi-channel (up to 8) autonomous audio gathering devices with Edge AI processing.

2.2.1 XMC 2Go board

The first option, which provides off-the-shelf the smallest formfactor, is to combine the Shield2Go board with the directly-compatible XMC 2Go board, as shown in Figure 9.

The XMC 2Go Kit with XMC1100 is maybe the world's smallest, fully featured microcontroller evaluation kit, showcasing an XMC1100 (ARM® Cortex™-M0 based). It includes an on-board J-Link Lite Debugger, power over USB (Micro USB), 2x user LEDs, and a 2x8 pin header suitable for connecting additional sensors.

The XMC 2Go board can be used with the Arduino development environment. Open source code can be found on [1].

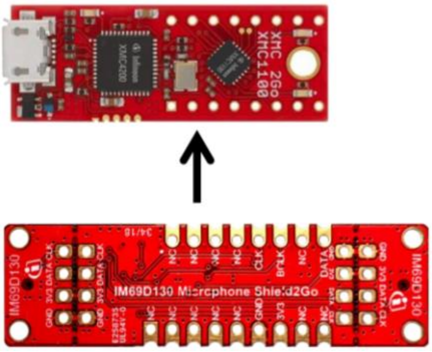


Figure 9. Shield2Go board and XMC 2Go

2.2.2 IFAG Audiohub – Nano

The Audiohub – Nano, depicted in Figure 10, provides the following features:

- Audio streaming over USB interface
- Powered through Micro USB
- 48 kHz sampling rate
- 24-bit audio data (stereo)
- Normal mode and low power mode with four pre-defined gain configurations
- Button to select mode and gain configuration
- LED indication for the configured gain level in normal mode and low power mode
- Volume unit meter display with onboard LEDs
- Mono and stereo mode
- External PDM connector

[1] https://github.com/Infineon/IM69D130-Microphone-Shield2Go/

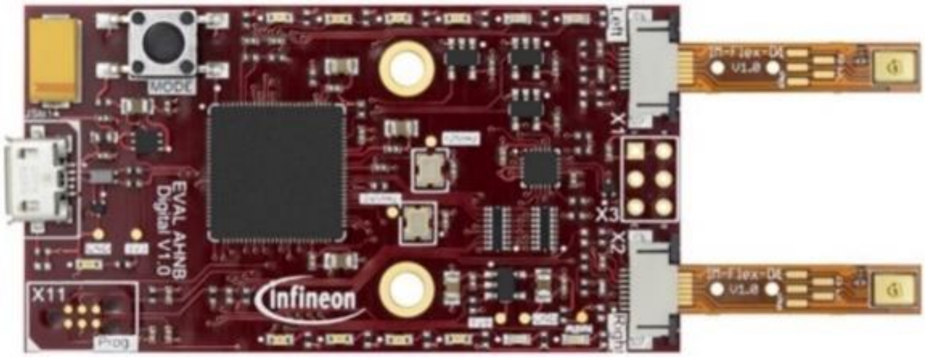


Figure 10. IFAG Audiohub – Nano

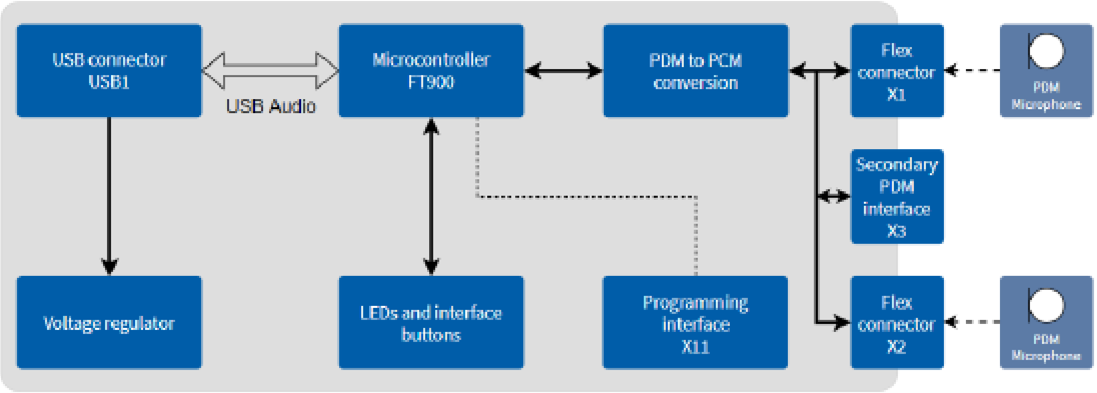


Figure 11. IFAG Audiohub - Nano block diagram

The board does all the required processing to transform the PDM input from the two microphones to two channels USB Audio. This allows for an easy integration in edge nodes featuring a USB input like, for example, a Raspberry Pi, without requiring low-level programming, since standard OS, like most Linux distributions and Windows releases, already feature USB Audio drivers off-the-shelf. No additional software installation is required, as the board is directly recognised as a native microphone.

2.2.3 IFAG Audiohub - Nano 4 Mic Version

In the scope of the MARVEL project, IFAG developed a version of the Audiohub – Nano with 4 of the IM69D130 microphones, as shown in Figure 12. All other features are similar to the Audiohub – Nano 2 microphones. It is also natively recognised as a USB microphone, in this case with four instead of two channels. The two additional channels enable more complex audio data processing techniques, as those envisioned in the MARVEL project.

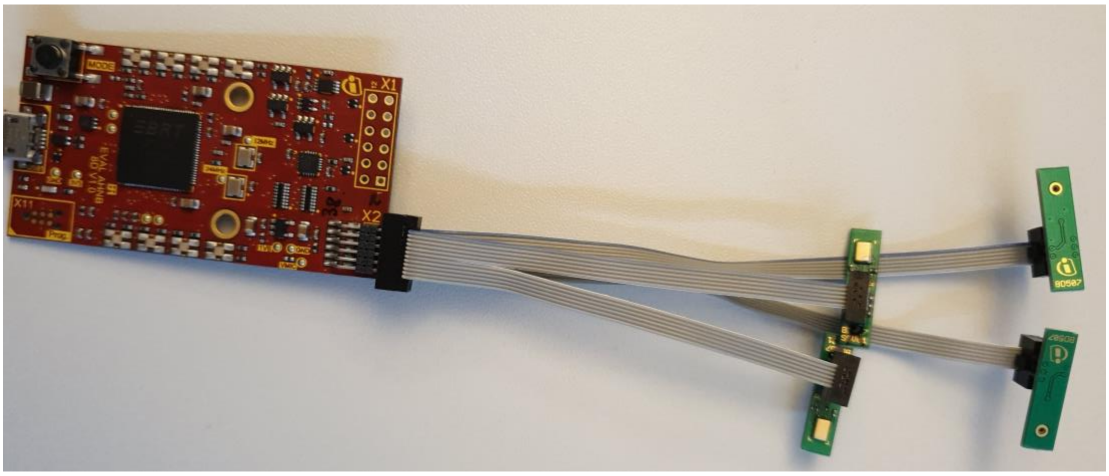


Figure 12. Audiohub - Nano 4 microphones

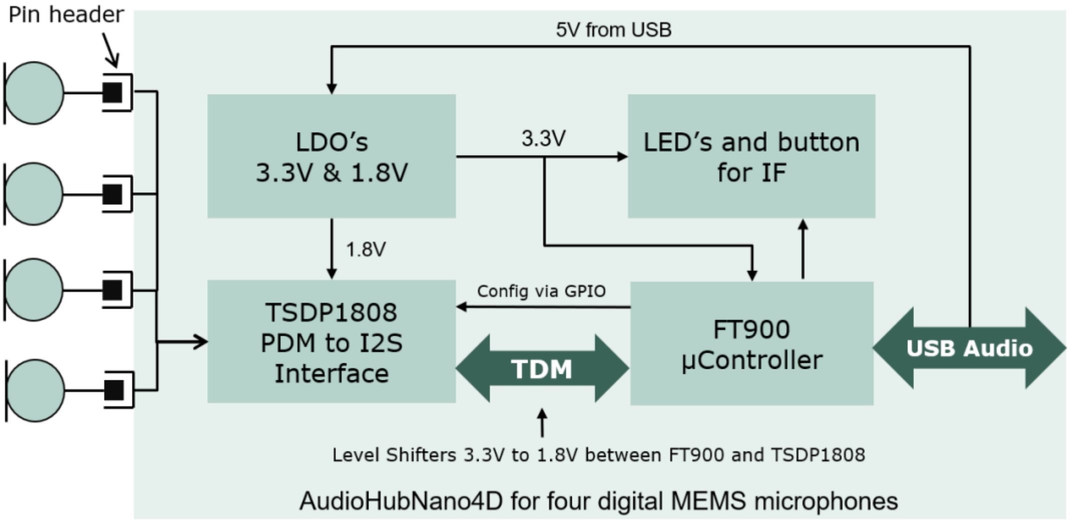


Figure 13. Block diagram Audiohub - Nano 4 microphones

2.2.4 Customised XMOS XK-USB-MIC-UF216

An over-the-counter microphone board and processor[2] was customised for ultra-low latency applications. Instead of the original 7 microphones that the original board featured, in the customised setup, 4 microphones were controlled by the processor, using a custom firmware designed for ultra-low latency.

- Default features:
  - 7 microphones, extendable up to 32
  - Audio streaming via USB
  - Sample rate 48 kHz
  - Dynamic range up to 100 dB
- Customised adjustments:
  - Up to 4 microphones
  - Ultra-low latency (~ 5x original speed)

[2] https://www.xmos.ai/microphone-aggregation/

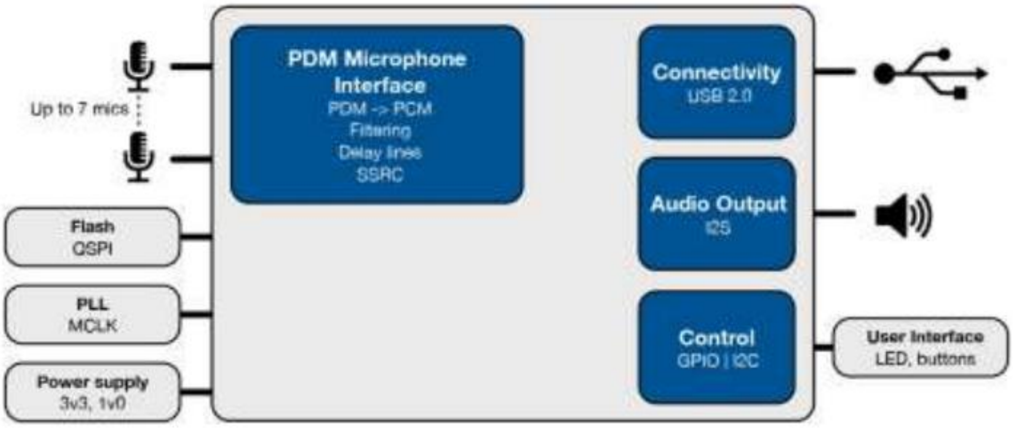


Figure 14. XMOS XK-USB-MIC-UF216 architecture

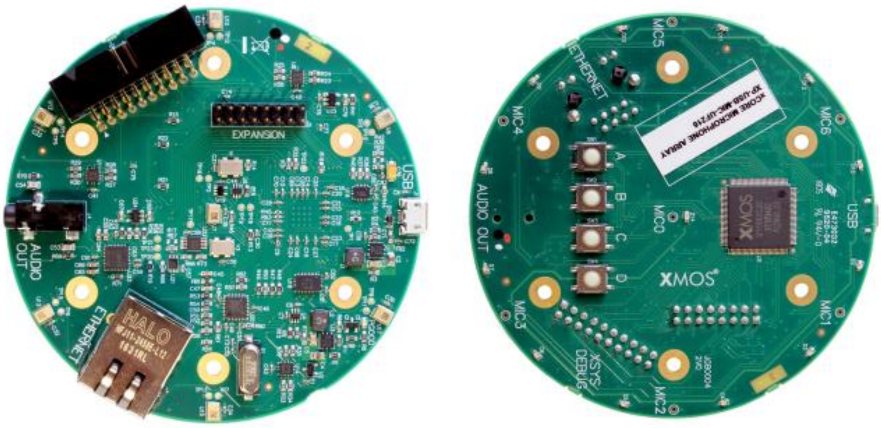


Figure 15. XMOS XK-USB-MIC-UF216 original PCB

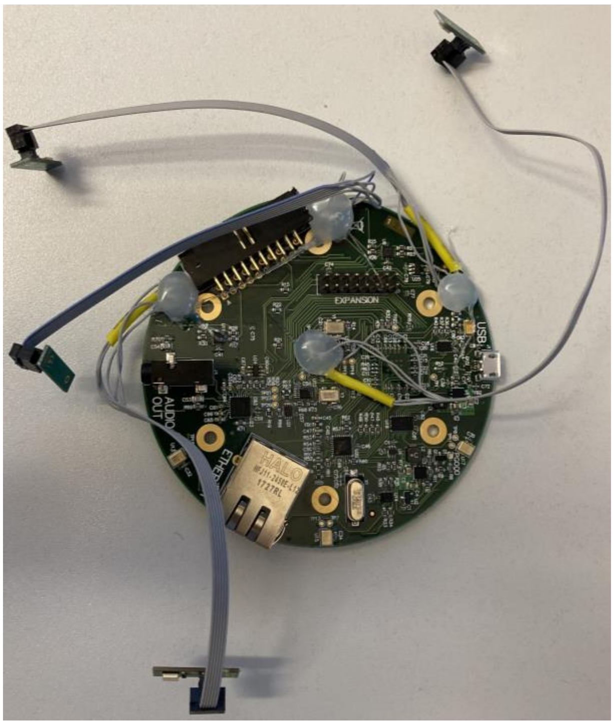


Figure 16. Customised XMOS XK-USB-MIC-UF216 with 4 IM69D130 microphones

2.2.5 Outlook: IFAG Audiohub – Nano 8 Mic Version

IFAG is currently working on an 8-microphone version of the Audiohub – Nano (Figure 17). It is a dual-PCB stacked design. It will consist of a separated microphone board and a main board, which is used for data processing. Figure 18 shows the layout of the microphone PCB. The microphones are arranged in a circular pattern to allow for optimal direction finding. Since the microphone board is connected via a header, it is very flexible, and different microphone configurations and geometries can be evaluated by developing additional boards.

It has the following features:

- Power source: USB connector
- Audio Data (up to 8 channels) streaming via WiFi to an external Cloud
- WiFi Module (Murata 1DX WiFi-Module)
- PSoC 64 processing board (Figure 20)

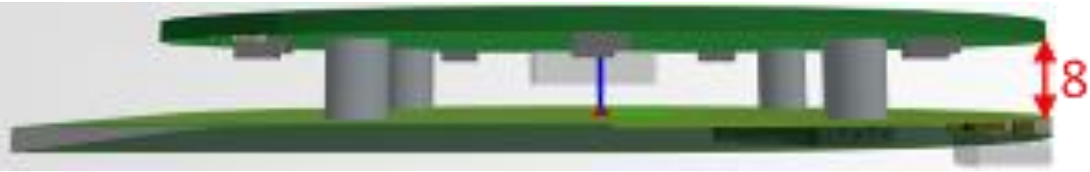


Figure 17. Dual-PCB 8-microphones Audiohub – Nano

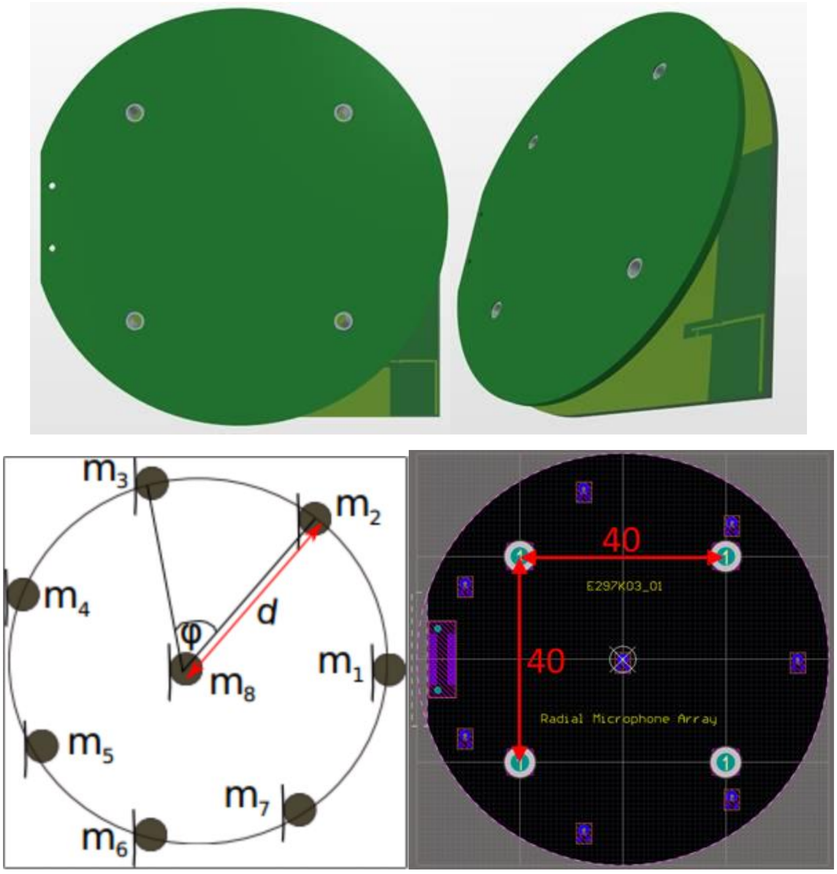


Figure 18. Layout of the microphone PCB

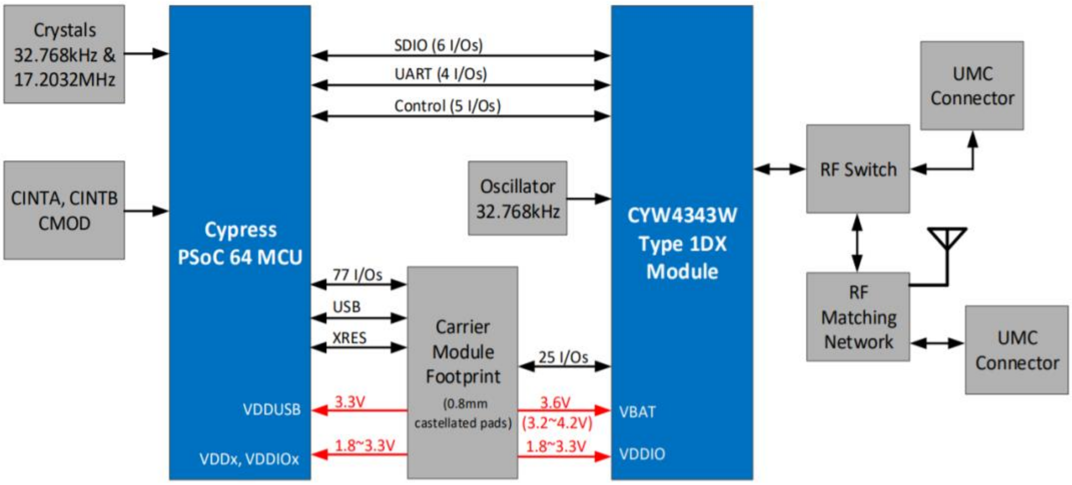


Figure 19. Block diagram processing board of Audiohub - Nano 8 microphones

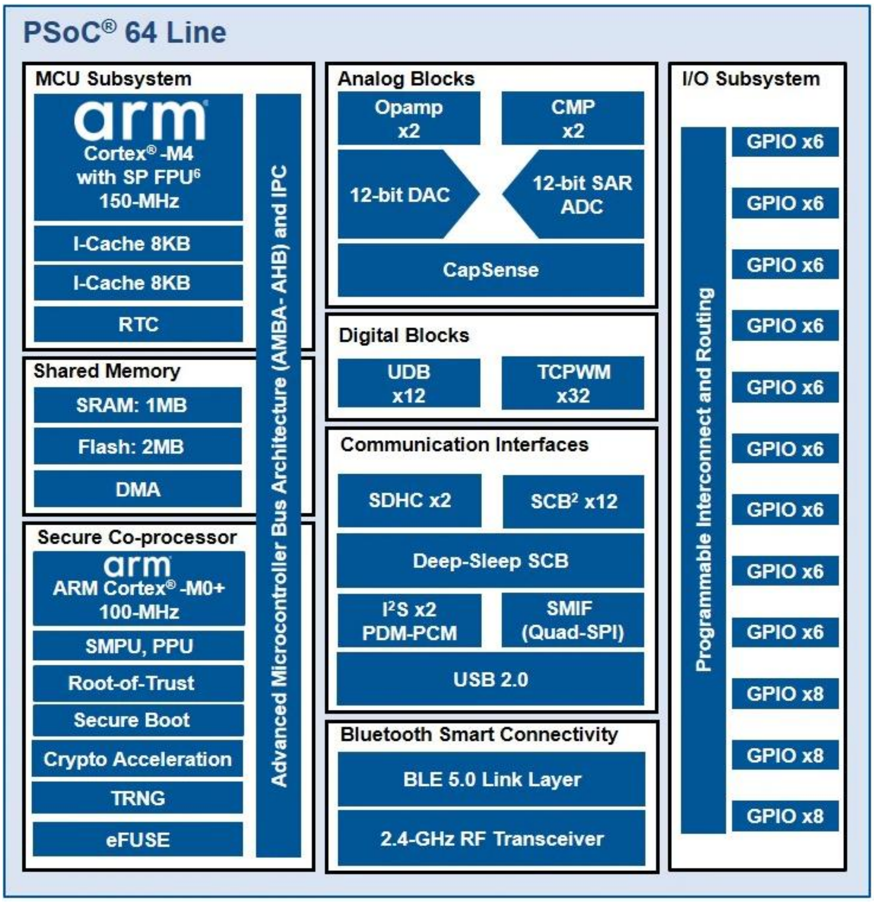


Figure 20. PSoC 64 block diagram

The PSoC 64 is the processing unit of the Audiohub – Nano with 8 microphones. Figure 20 shows its capabilities. It will be delivered with a default firmware streaming the raw 8-channel audio data via WiFi to an external cloud but can be programmed and debugged for extended functionality via JTAG or SWD.

The software ModusToolbox™ Machine Learning, publicly available online at[3], allows for rapid evaluation and deployment of ML models on all IFAG MCUs, including the PSoC 64 (Figure 21). A customised firmware can be developed to on-board Edge AI processing, streaming higher-level data to the cloud for traffic reduction, potentially improved latency and privacy, cloud offloading, and improved resiliency and autonomy.

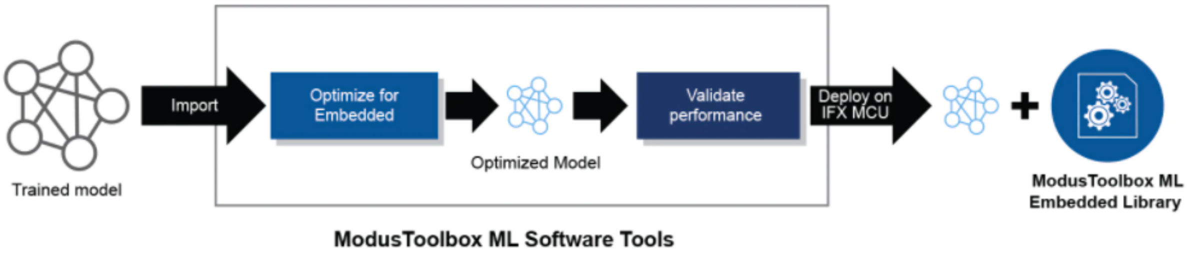


Figure 21. Flow for ML on the PSoC 64 with ModusToolbox

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

2.3 Early edge devices experimental results

FBK experimented with the 2 microphones version of the Audiohub - Nano. The board was connected via USB to a Raspberry Pi device and to a laptop. The former set up was used to implement the data stream from the Raspberry Pi edge device to the fog layer, towards the microphone deployment in the MT pilot. The latter configuration was used to perform the test recordings towards the collection of the staged dataset. Figure 22 shows the Audiohub - Nano board connected to the Raspberry Pi via USB.

At the moment, no issues were identified in terms of quality of the signals. Note that the default gain is 0 when the device is turned on, but signals can be amplified afterwards without any quality deterioration.

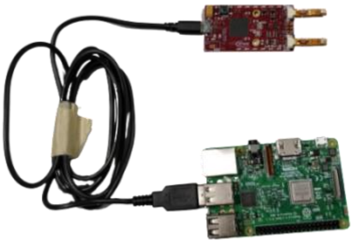


Figure 22. Mems microphones connected with the Raspberry Pi

As part of the UNS Drone experiment use case, collection of audio recordings was performed. The recordings were made both on the ground and on the drone. The recording equipment was provided by IFAG and consisted of 4 sets of micro - USB enabled development board Audiohub - Nano each with 2 MEMS microphone (stereo configuration).

The initial step in the process was to determine which device and which application will be used to start the recording on the board. We decided to use mobile phones, since the usage of laptops would be impractical for the recording in public spaces and mounting laptop to the drone is also not possible. We also considered using other devices such as Raspberry Pi, but their usage is also not so convenient since external battery is needed. The recording app should meet the following two criteria:

1. Supports recording from USB interface
2. Supports scheduled recording

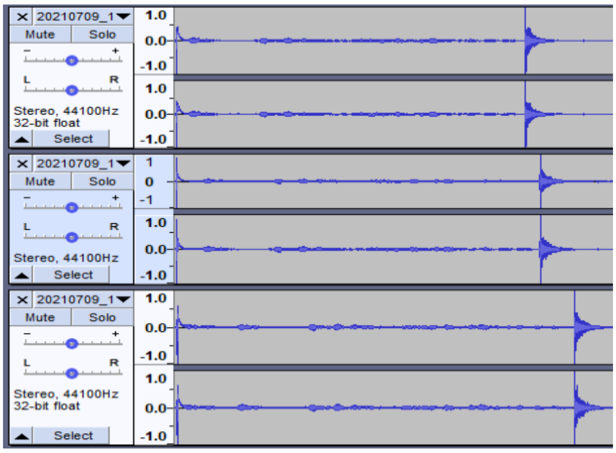


Figure 23. Time shift in signals recorded by different phones

Initially, we tested the SensMiner app provided by AUD. At the time, this app did not support scheduled recording, therefore we tried to find an alternative solution. During the process of testing a number of apps available on Google Play, we found that an app named RecForge II4 fulfills both criteria.

When we found the appropriate recording app, a number of experiments were conducted in closed space in order to determine the best possible configuration of the board. Based on the magnitude levels of the recording sound we determined that it is optimal to use the normal power mode with the highest gain of 24 dB.

During these experiments, we also concluded that synchronisation between files recorded on the different phones was not perfect, as shown in Figure 23. and that some reference signal should be used. Specifically, the time shift in the position of reference sounds is not caused by different distances between source and microphones. The largest distance between two microphones was 3.5m which makes the possible time shift of approximately 10 milliseconds. The observed shifts on the other hand were even larger than 0.5 seconds. These shifts were caused by the fact that recording on some phones could have started earlier than in the others.

The second group of experiments was conducted in open space, specifically in the main square of the campus of the University of Novi Sad. The exact positions of the recording points in these experiments are given in Figure 24. At each of the indicated positions (M1-M4), one Audiohub - Nano board with two microphones was mounted and connected to a cell phone via USB cable.

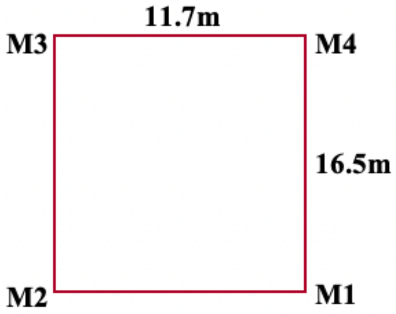


Figure 24. Positions of microphones in open space experiments

The experiments in the open space could be classified in the following three groups:

1. **Exp. 1 – linear movement:** The sound source reproducing speech is moving in M1-M4 direction starting from M4 and moving from both M1 and M4.
2. **Exp. 2 – circular movement:** The sound source reproducing speech is moving on the edges of the rectangle formed by the positions of all recording microphones.
3. **Exp. 3 – whistle sound:** The whistle sound reproduced in all corners of the rectangle formed by the positions of all recording microphones.

From the analysis of the files recorded in Experiments 1 and 2, it can be concluded that human speech is intelligible whenever individuals are 10 meters apart or less from the microphones. Experiment 3 shows that reference signal can be successfully captured at a position more than 20 meters away from its position.

After the test experiments outlined above, UNS has performed the first staged recording. The staged recording was performed using the same configuration shown in Figure 24, with the

[4] https://play.google.com/store/apps/details?id=dje073.android.modernrecforge&hl=hr&gl=US

exclusion of position M2, where we experienced technical problems with the recording setup. Microphones were set at a height of approximately 1.7m and attached to public light poles as shown in Figure 25. The individuals in these experiments were placed in the upper part of the recording scene presented in Figure 24 (closer to microphones M3 and M4). In all recordings, background noise produced by the drone is present. Intelligible speech sound was captured by the microphones in positions M3 and M4, while recordings obtained in position M1 mainly contain drone noise.

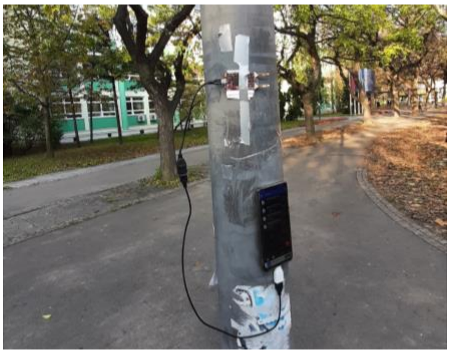


Figure 25. Recording setup: Audiohub - Nano with two microphones connected by USB connection to a mobile phone

The Audiohub - Nano microphone module was tested on the Portable GRNEdge camera setup. This setup included a Raspberry Pi 3 Model B+, a Pi cam for video and the Audiohub - Nano microphone as the acoustic sensor. This version of GRNEdge was enclosed in a weatherproof plastic box. Since the Audiohub - Nano is not weatherproof, it had to be installed inside the sealed container with the other components. A custom casing was designed, and 3D printed to fix the Audiohub - Nano inside the sealed container and prevent it from moving. The mount is shown in Figure 26.

This setup muffled the sound being collected and the sound level picked up by the microphones was very low. Thus, the gain of the sound signal was amplified with software. However, the quality was low due to clipping and low SNR. Ideally, this problem is mitigated with the use of a better physical interface between the microphones and the outside world, for example, using a separate waterproof case for the Audiohub - Nano and microphones.

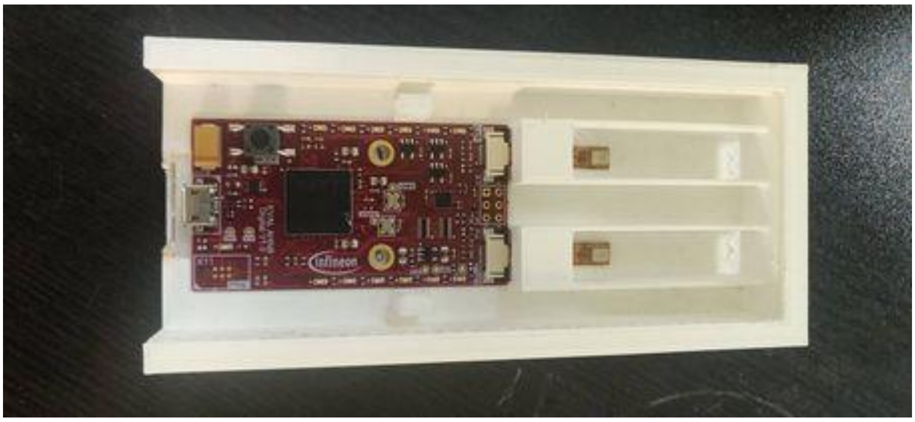


Figure 26. 3D printed mount for Audiohub - Nano

2.4 Pre-processing for Data analytics, as a basis for light-weight ML models

Components performing visual and audio-visual data analysis in MARVEL need to be deployed as close to the edge as possible, in order to allow for decentralised (fast) inference, decreased data transfer, and to better satisfy anonymisation requirements. However, the high computational and memory requirements associated with processing visual data streams generate challenges. These components need to be deployed on embedded GPU platforms (e.g., NVIDIA Jetson TX2 or XAVIER modules) installed close to the edge, or fog. To perform real-time inference, they need to be connected on light-weight backbone networks, and optionally provide a just-in-time inference capability, i.e., when real-time operation is needed and the computational resources are not sufficient to compute the output of the DL model, an early exiting can be used to provide a prediction which can be less accurate. The MARVEL visual and audio-visual crowd counting components will provide such an option.

Another important aspect that needs to be considered in the design and implementation of the data analytics capabilities in the MARVEL framework is related to the resource allocation of the different components deployed at the edge, the fog, or cloud sites. Simultaneous deployment of multiple data analytics components on the same platform requires increased computational power and memory consumption which, in the case of limited edge or fog computing platform capabilities, will lead to suboptimal operation. The adequate operation frequency for each component will need to be determined towards addressing this challenge.

The deployment of audio and audio-visual inference models at the edge has to consider the constraints imposed by the limited resources available at the edge devices and the operational requirements imposed by the specific use cases. Precisely, even considering powerful edge devices equipped with GPUs, running multiple instances of AI models, might become a computationally intensive task whose complexity might increase, especially if one considers that inferences should run under (near) real-time constraints. To enable AI model to run on edge and resource-constrained devices it becomes of paramount importance to find ways for limiting their computational complexity. Viable solutions to achieve such an objective are (i) to compress large AI models by reducing their number of parameters limiting the (possibly) consequent accuracy degradation (ii) to design light-weight AI models able to achieve produce both accurate predictions within the temporal constraints imposed by the specific use case at hand. These aspects will be investigated and developed within the MARVEL framework and exploited for performing optimal audio-visual data analysis.

The use of edge devices for processing and analysis of sensor data allows better protection and privacy, and reduces the amount of data to be transferred. Sensor data such as audio, images, and video, contain sensitive personal information, thus transferring such data from sensors to cloud processing and analysis creates risks that personal information will be leaked and misused, even when secure transfer protocols are used. By processing and analysing data in the sensor, i.e., on the edge, allows extracting and transmitting only such information from sensor data that is relevant to applications, so that sensitive personal information does not need to be transferred. This could include, for example, using ML models that will recognise target classes of sounds or images and transmit information only about the recognised classes. Thus, edge processing provides a way to handle what information is being sent to the cloud, and what information is discarded. Furthermore, transmitting only information about the recognised events takes significantly less bandwidth in comparison to transmitting raw audio, image, or video data.

MARVEL’s E2F2C testbed consolidates a heterogeneous and disperse hardware environment under a common container-based orchestration framework implemented by Kubernetes – at least at the Cloud and Fog layers. Additionally, it provides the necessary tools and utilities to easily connect and manage the devices connected at the Edge. Work performed in T3.4 decides and optimises the placement of both local and distributed AI tasks at one or more layers of the infrastructure, including the Edge. Therefore, as part of the adaptive E2F2C distribution and optimisation of AI tasks, specific light-weight ML functions may be sent for execution closer to the sensors, enabling early decisions near the data source with low latency. Moreover, the deployment technology from T3.4 will assist in the collection and early processing of audio landscape data, particularly from MEMS devices installed at the Edge, by enabling the pre-processing of data using edge ML and pattern recognition algorithms.

In the MT application scenarios, edge processing of the audio and visual streams captured by the sensors serves two purposes. First of all, edge processing addresses privacy issues by avoiding the transmission and storage of audio-visual data of the citizens in public spaces. This limits the risks related to misuse or leak of sensitive personal data. The second purpose is to reduce the amount of bandwidth and related energy needed to transfer the data from the sensors to the data-centre, for processing or storage. At the moment, due to the issues just mentioned, only low resolution and low frame-rate videos are recorded from the cameras.

3 Intelligent audio data analysis and collection

T4.2 is composed of two principal parts: intelligent audio analysis and privacy compliance. These two subtasks are strongly dependent, where the audio analysis will only take place once the right privacy measures have taken place. Such measures are the identification and complete removal of speech segments; therefore, speaker information is discarded and the audio content can be used properly for further analysis whether by AUD or MARVEL partners, without any privacy violation. However, it is important that this pre-processing step of speech identification and removal takes place on edge devices. This is in order to ensure that no sensitive speech data gets transferred from edge devices (e.g., microphones) to fog or cloud nodes, and therefore maximum privacy respect is guaranteed. This pre-processing step can be achieved with the integration of AUD's devAIce SDK, which contains the Voice Activity Detection (VAD) module capable of detecting speech segments in real-time. devAIce is expected to run not only on powerful CPU/GPU machines (cloud/fog) but also on medium-high end edge devices (e.g., Raspberry Pi). Thus, as a first step, devAIce will be deployed on edge devices (MEMS microphones) with respect to T4.1, where the pre-processing will take place. For this particular step, devAIce's VAD model has been updated to achieve better performances.

The VAD model has also been upgraded to detect not only speech segments detection but also overlapping music segments. Such a feature will be valuable for pilots, either for analysis or for data collection.

Another important step is the audio data collection for either training or in-domain benchmarking. For this, AUD's SensMiner toolkit has been updated accordingly to adapt to the UNS pilot use case.

3.1 DevAIce platform

devAIce is AUD's modular technology for intelligent audio analytics, including the award-winning openSMILE audio feature extraction toolkit, and modules for sound event detection, acoustic scene classification, and speech analysis. devAIce is written in C++ and is configured to function on powerful computing nodes as well as medium and high-end edge devices (Raspberry Pi). A trimmed-down version of the toolkit containing just the openSMILE toolkit also exists and can be used on edge devices with limited computational capabilities.

devAIce comes with an interface in Python, iOS, Android and C. This doesn't mean that it cannot be used with other programming languages, however, in such scenario, the user has to manually build his own wrapper around the C interface, as the Python, iOS and Android ones are simply wrappers built around the principal C interface. This portability and flexibility of deployment makes its use straightforward for multiple scenarios. As mentioned above, almost all the modules of the toolkit can be deployed and used on high-end edge devices. This applies to the VAD module.

With this portability property, the VAD module is to be deployed on microphones with a high-end integrated processor, in order to identify speech segments and remove them before propagating the audio data to the fog and cloud nodes. Moreover, any audio analytics should preferably take place on premise, and if resources do not allow it, the analytics and models inferences will be executed on the fog or cloud layer, but taking as input the pre-processed audio sequences, from which speech has been detected and removed. This way, speaker information is discarded and privacy is maintained and respected. A different way of using the VAD module for privacy compliance, instead of removing speech segments, is to apply voice anonymisation techniques on the detected speech segments. These techniques can be signal processing based such as applying low-pass filtering, as most of the voice content is localised to high frequencies, it is possible to low-pass filter the detected speech segments at 250 Hz, by sub-sampling the signal at 500 Hz and up-sampling the results back to 16 kHz. There have also been proposed other more complex techniques, such as MFCC inversion for voice anonymisation [1]. Figure 27 shows an overview of the audio analytics pipeline inside MARVEL architecture where raw audio will be pre-processed by devAIce to detect speech segments. Those speech segments will be either discarded or subject to blurring techniques, and then the speaker information free audio will be used by DL model for predictions.

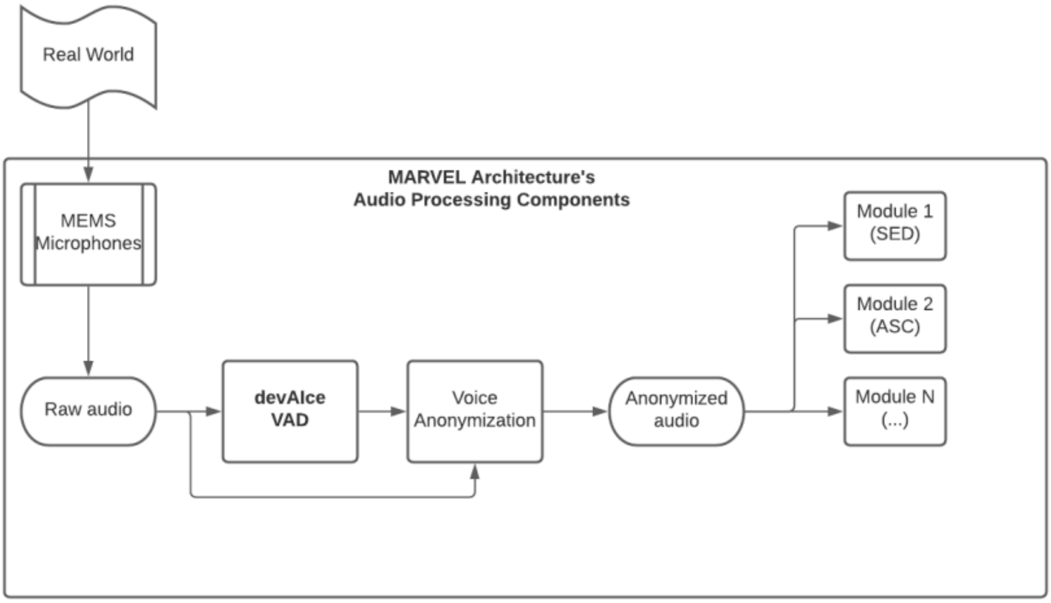


Figure 27. Overview of the audio analytics pipeline

3.1.1 Features extraction via openSMILE

The openSMILE toolkit is a state-of-the-art feature extraction tool for audio streams. It supports several standard feature extraction algorithms (e.g., spectrograms, MFCCs, etc.). The specific features and hyperparameters (e.g., window size for spectrogram) thereof are configured by the user in a configuration file. openSMILE accepts a raw audio stream as input (uncompressed PCM in either 32-bit float or 16-bit signed integer format) and returns a stream of features conforming to the configuration specifications. openSMILE works locally and does not stream any data or metadata to AUD’s infrastructure.

Several feature extraction algorithms supported by openSMILE are not privacy-preserving. For example, spectrograms contain information both about speaker identity and speech content, thus infringing upon the privacy of individuals. Privacy considerations need to be handled outside of openSMILE, e.g., either by anonymising the audio streams beforehand, or anonymising the audio features afterwards. If openSMILE, as well as the model’s inferences, will be executed on premise, there will be no need to apply anonymisation techniques, unless these features will be propagated to higher layers (fog or cloud).

openSMILE will be of use on limited computational edge devices, where resource-hungry DL models won’t be able to be deployed, and therefore the need for less resource-hungry ML architectures. These ML architectures (e.g., SVN) are not expected to work well with frame-level features such as spectrograms or MFCCs. However, openSMILE can extract a set of features that are calculated on clip-level (over the entire audio sequence) by applying high-level statistics on these low-level features. This kind of high-level features can be fed to ML models for inferences and predictions. It is important to note that this is an extreme case, which will be referred to only if there is no possible way to quantise DL models enough to be supported on the limited computational edge devices, as performance gap will be noticeable if we refer to the use of simpler architecture rather than DL models for certain tasks.

The openSMILE toolkit is a feature extraction tool for audio streams. As such, leveraging the addition of audio features, coming from openSMILE, depicting events can lead to meaningful insights regarding the detected events in the decision-making platform. SmartViz will constitute MARVEL's project UI and has the capability of showcasing audio clips referring to the detected events it visualises. By configuring the options of the data input mechanisms feeding SmartViz and with the assistance of other components that are included in MARVEL workflows, the Decision-Making Toolkit (DMT) will be able to give the users access to a detailed and global outlook of detected events. The audio clips combined with visual references from other components of MARVEL can be used as a quick validation of the insight's accuracy that MARVEL provides to the users via the DMT and better support their decisions.

3.1.2 VAD toolkit

The VAD model has been updated and redeveloped according to a novel state-of-the-art architecture. The architecture is inspired from the paper published by Lee et al. [2], in which a novel concept is proposed. A new type of attention module is introduced and these modules are based on convolutional layers. Also, the attention will not be calculated on time-domain only, but also on the frequency domain; this concept is referred to as "dual attention". The attention modules can be visualised as depicted in Figure 28, where H is the hidden state of dimension T (sequence length) and D (number of hidden features). Dual attention is the combination of both attention modules.

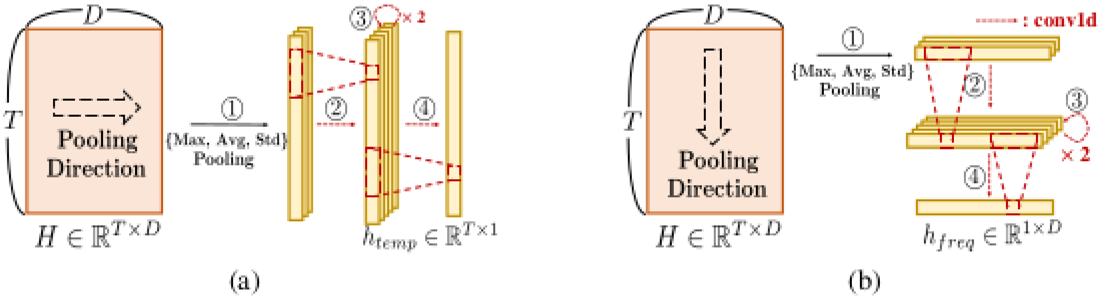

D
T
Pooling Direction
$H \in \mathbb{R}^{T \times D}$
①
{Max, Avg, Std} Pooling
②
③
× 2
④
$h_{temp} \in \mathbb{R}^{T \times 1}$
(a)
: conv1d
$h_{freq} \in \mathbb{R}^{1 \times D}$
(b)

Figure 28. (a) Temporal attention (b) Frequential attention

With these two attention modules integrated, an LSTM (Long short-term memory) model has been built. Figure 29 shows an overview of the architecture.

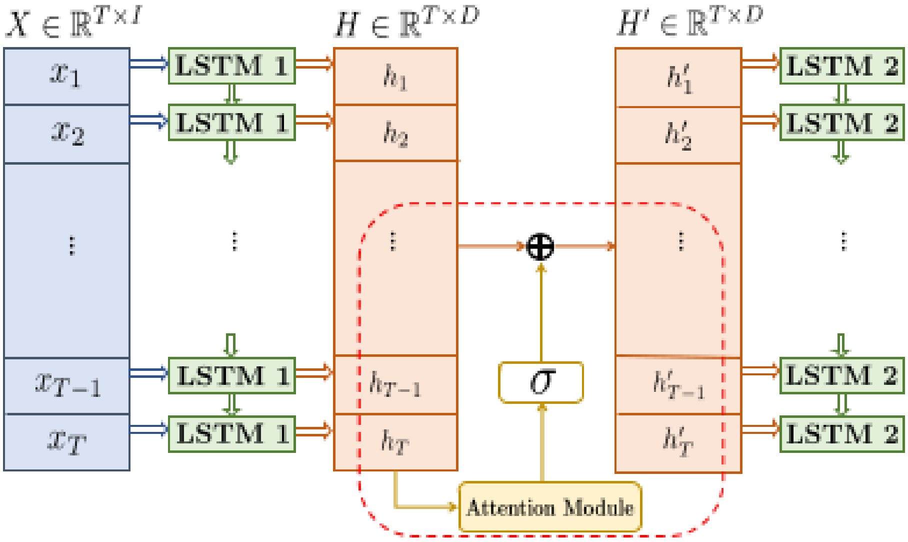


Figure 29. VAD model's global architecture

The model has been trained on artificially mixed data. This data has been generated by mixing multiple internal databases of sound events, speech, and background noises, leading to generate audio data with properties similar to data recorded in the wild. The framework developed for data mixing is flexible and reproducible, which means, when in-domain data will be available from the partners, it can be included in the mixing process in order to generate more accurate in-domain data that can be used for re-training the model.

The model has been tested on the artificially mixed data as well as other databases such as MUSAN [3]. With this novel state-of-the-art architecture, the UAR has increased from 79.8% to 83.8%.

The new model integration into devAIce is in process. In the meanwhile, resources consumption tests were conducted. The new model consumes nearly 20MB for inference. It also exploits multi-threading which makes it 4 times faster on a multi-core processor, but as the architecture is more complex, it is slightly slower on single-core processors. Table 4 shows this behavior in terms of numbers.

Table 4: Time elapsed on single and multi-core processors

	Time elapsed on an input of 1449s (multi-core)	Time elapsed on the same input (single-core)
Old VAD model	5.44s	**5.42s**
New VAD model	**1.37s**	6.53s

With this behaviour, it is expected that the new model will be faster on powerful computational nodes (fog and cloud), and slightly slower on limited edge devices.

With respect to T4.1, the next step when devAIce will be delivered to FBK is to deploy it on the developed devices based on MEMS microphones. These devices are known to contain high-end processors (e.g., Raspberry Pi) that will allow the full use of devAIce on premise.

3.1.3 Music Detection

With this new state-of-the-art architecture, the model has been trained in a way to detect overlapping music and speech, thus the model can be used as a voice activity detector, as a music detector or both. The training data is also generated with the data mixing framework recently developed; music databases were included this time in the mixing process. The music detection performance has been evaluated on the artificially mixed dataset and MUSAN as illustrated by the confusion matrixes in Figure 30.

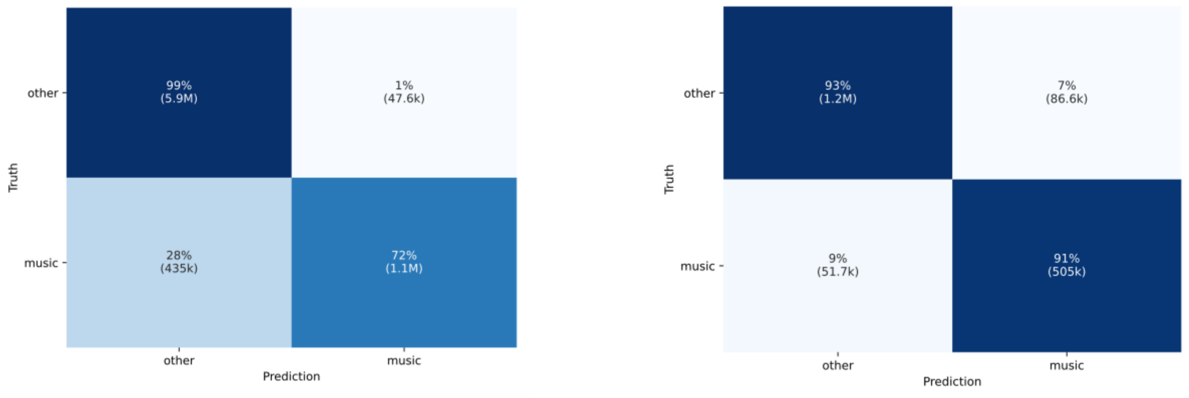


Figure 30. Confusion matrixes for music detection. Left: Artificially mixed dataset (34 hours), Right: MUSAN (10 hours)

As the model that detects voice activity detection and music is the same, which will be referred to as mutli-head (multi-task) model, integrating the new VAD model in devAIce will lead to also having the music detection feature available.

This music detection feature will be mostly useful for the UNS pilot where a staged university party will be monitored. It is definitely interesting to be able to differentiate between a potential fight and violent activity or simply a group of youngsters playing very loud music on speakers. An example use case is when high volume sound is detected. This music detection model will be used to confirm if it is music. In that case, there is nothing to worry about, otherwise, further investigation will be required as potential risk is observed.

The music detection can be also used by the MT pilot, as the same scenario explained above can also happen in a large square in a city. Of course, these are not the only scenarios and the partners are invited to use this music detector wherever they consider to be useful.

3.2 SensMiner toolkit

SensMiner is an Android app developed by AUD to record environmental acoustics as well as user annotations. While the audio is being recorded, the user can in parallel annotate it and store the corresponding segment in the phone memory. SensMiner collects raw audio, GPS information, and user tags. Audio is recorded in 16bit PCM format at 44.1 kHz. All data is stored as JSON files on the user's smartphone and need to be manually transferred.

It is a standalone app used exclusively for data acquisition by the UNS pilot and it does not perform any processing or analytics. To adhere and work well with the UNS use case, SensMiner has been updated to support the latest Android version as well as to include scheduled recordings.

As stated in Section 2.1, the initial stage recordings within the UNS drone experiment use case were carried out without SensMiner, as the app was not available at the time. UNS received an updated version of the app after the initial staged recording and will use it in the next recording sessions. In the meantime, initial testing was performed, and we report here the main conclusions. It can be confirmed that the app now supports USB recording and thus it can be used with IFAGs AudioHub - Nano. In addition, it also supports scheduled recordings which is important for the UNS Drone experiment use case. Some further improvements, refinements and such that are relevant for this use case, possibly for future work, are the following:

- The settings screen is currently inaccessible (i.e., the change of recording settings is not possible).
- The recorded files should be reproducible from the app.
- The recorded files should be easily accessible; currently, files can be accessed only when the phone is connected to the PC and Android Studio must be used.

4 Conclusions

In Task 4.1, different platforms featuring the selected high-end MEMS microphone have been successfully evaluated. Relevant project partners already have valuable hands-on experience with the MEMS devices, being able to record and analyse sound, conveniently via USB, up to 2 (stereo) synchronised channels. In the second half of T4.1, platforms with more channels (up to 8 synchronised audio channels) and features (Edge AI capabilities) will be developed and distributed for evaluation and fulfillment of the MARVEL use cases.

Task 4.2 can be divided into two principal aspects, i.e., intelligent audio analysis and voice anonymisation. Progress has been achieved on both aspects. devAIce‘s VAD module, which is critical and necessary for voice anonymisation, as audio analysis will only be carried out after detection and filtering speech from the audio stream, has been updated according to a novel state-of-the-art approach and is currently in the process of getting integrated in the SDK. The same model has been upgraded to detect overlapping music, an interesting feature that can be exploited by all pilots. In addition to that, SensMiner has been updated to adhere to the UNS pilot requirements for data collection. To achieve this the application now supports the latest Android version as well as scheduled recordings, a critical feature for the pilot.

5 References

[1] Cohen-Hadria, A., Cartwright, M., McFee, B. and Bello, J.P., 2019, October. Voice anonymization in urban sound recordings. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.

[2] Lee, J., Jung, Y. and Kim, H., 2020. Dual attention in time and frequency domain for voice activity detection. *INTERSPEECH 2020*, (pp. 3670-3674).

[3] Snyder, D., Chen, G. and Povey, D., 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484.*