European Commission | Horizon 2020 European Union funding for Research & Innovation

Big Data technologies and extreme-scale analytics

Multimodal Extreme Scale Data Analytics for Smart Cities Environments

D4.4: Optimal audio-visual capturing, analysis and voice anonymisation[†]

Abstract: This report focuses on the final version of the optimal audio-visual capturing, analysis and voice anonymisation. With regard to optimal audio capturing, the first section of the deliverable offers a detailed description of the development progress of the hardware used within the scope of this project, since the initial version of the document. Moreover, the applied methodologies to provide pre-processing and data analysis through the usage of Edge AI techniques are presented. The last part describes the deployment process of the final version of the hardware within the different project use cases as well as the final experimental results. The second section presents the progress made with regard to intelligent audio analysis with the Voice Activity Detection module within the scope of devAIce SDK, the modular technology encapsulating multiple AI modules designed for cross-platforms deployment, as well as the upgrades applied to the data collection and user annotations toolkits, SensMiner and the AVER application. A third section in this document describes the toolkit used to anonymise the sensitive speaker information, AudioAnony, as well as the approach adopted to design and build the voice anonymisation pipeline on the edge along with a detailed description of the deployment process along the different use cases. Furthermore, the progress of the KPIs set for the tasks and the relevant components is discussed in a separate section.

[†] *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	31/12/2022
Actual Date of Delivery	30/12/2022
Deliverable Security Class	Public
Editor	*Bahaeddine Abrougui (AUD)*
Contributors	AUD, FBK, CNR, IFAG, UNS, ZELUS
Quality Assurance	*Borja Saez (IFAG)* *Stella Markopoulou (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. *Borja Saez (IFAG)*
2. *Stella Markopoulou (ZELUS)*

Revisions

Version	Date	By	Overview
v1.0	30/12/2022	Editor, PC & Contributors	Review comments from the PC addressed – Ready for submission
v0.9	15/12/2022	Editor & Reviewers	Merged second phase of review from ZELUS and IFAG – Ready for review from the PC
v0.8	06/12/2022	Editor & Contributors	Second round of input complete – Ready for second review
v0.7	22/11/2022	Editor & Reviewers	Merged review from ZELUS and IFAG – Ready for input round 2
v0.6	11/11/2022	Editor & Contributors	First round of input complete – Ready for review
v0.5	24/10/2022	Editor	Updated TOC – Ready for input
v0.4	12/10/2022	Editor	Updated TOC – Ready for review
v0.3	11/10/2022	Editor	Updated TOC
v0.2	06/10/2022	Editor	Updated TOC – Updated reviewers
v0.1	30/09/2022	Editor	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.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AVER	Audio-Visual Emotion Recognition
DMT	Decision-Making Toolkit
E2F2C	Edge to Fog to Cloud
EC	European Commission
FP	False Positives
JSON	JavaScript Object Notation
JTAG	Joint Test Action Group
I2S	Inter-IC Sound
IoT	Internet of Things
GPS	Global Positioning System
KPI	Key Performance Indicator
LSTM	Long Short-Term Memory
ML	Machine Learning
MQTT	MQ Telemetry Transport
PCM	Pulse Code Modulation
RMS	Root Mean Square
RNN	Recurrent Neural Network
RTSP	Real-Time Streaming Protocol
SDK	Software Development Kit
SQL	Structured Query Language
SOTA	State Of The Art
SSH	Secure Shell
SWD	Serial Wire Debug
UAR	Unweighted Average Recall
UI	User Interface
USB	Universal Serial Bus
VAD	Voice Activity Detection
VPN	Virtual Private Network
WP	Work Package
ZIF	Zero Insertion Force

Executive Summary

This deliverable is the final version of the optimal audio-visual capturing, analysis and voice anonymisation. The initial version has been released under the deliverable D4.1[1] in M12. The document reports the full progress of Task 4.1 and Task 4.2 achieved up to M24. Both Tasks are ending on M24.

The final goals of Task 4.1 and Task 4.2 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:

- Audio capturing innovation by analogue and digital microphones based on MEMS technology allowing low-weight ML models edge pre-processing.
- Audio data and environmental acoustics collection through a UI-based user annotation with SensMiner and AVER toolkits.
- Modular AI technology by devAIce platform for audio analysis, and on-premise audio anonymisation with VAD-AudioAnony.

The first part of the document, Section 2, presents the results obtained during the development of Task 4.1 *Optimised audio capturing through MEMS devices*. This section describes the latest state-of-the-art MEMS microphones, and the technical motivation, use and relevance for this project. This section also describes all the audio boards that have been used during the MARVEL project, focusing on the Nao Hub acquisition board (both 4 and 8 microphones version) that has been developed under the scope of this task to accomplish the project requirements. Furthermore, the section continues with a description of the different Machine Learning (ML) algorithms developed for audio signal processing. Finally, the section concludes with the examples of the usage of the developed components in the different pilot use cases.

The second part of the document, Sections 3 and 4, covers the optimal audio data analysis, collection and voice anonymisation. Section 3 focuses on the adopted approach for intelligent audio analysis, and describes the updates within the devAIce platform, a modular SDK compatible with cross-platform environments which is optimised to run on commodity clusters as well as high-end edge devices. The devAIce platform exposes the openSMILE tool that provides large feature space extraction from raw audio, and it allows low-weight ML models to perform their designated tasks. The SDK also encapsulates multiple AI models and algorithms, among which we find the VAD module that has been subject to a major upgrade and continuous development and optimisation for the MARVEL use cases. Such development processes include the integration of a music detection function, improvement of noise robustness as well as responsiveness and reactivity. Furthermore, the section also presents the progress made to optimise the audio data collection and the UI-based user annotations, respectively with the upgrade and development of the SensMiner and AVER toolkits.

Section 4 describes the tool used for speech anonymisation as well as the integration of the tool inside the audio anonymisation pipeline along with VAD. It also discusses the deployment process of the composite component within the relevant project use cases. A detailed discussion on the progress and achievement of the relevant KPIs also takes place in Section 4.

[1] MARVEL D4.1: Optimal audio-visual capturing, analysis and voice anonymisation – initial version, 2021. https://zenodo.org/record/6821280#.Y2pUXezP0UE

D4.4 documents the audio data acquisition and annotation as well as the pre-processing and processing hardware and software along with sensitive data anonymisation. This leads to a final solution for an optimal audio data capturing, analysis and voice anonymisation pipeline. This pipeline operates in a way such as the ingested raw audio data is directly processed on the capturing devices themselves, anonymising the sensitive information, applying the relevant ML algorithms, and forwarding the anonymised audio stream to the rest of the E2F2C (Edge-to-Fog-to-Cloud) pipelines layers for further complex analysis along with the obtained ML results, which will be visualised via the intermediate of MARVEL decision-making toolkit (DMT), SmartViz[2].

With this completed, the amount of data being transmitted is reduced while the amount of the collected data for further improvements is increased, ending up with a solution that complies with privacy constraints optimised for low-resource consumption and real-time processing.

[2] SmartViz https://www.zelus.gr/smartviz/

1 Introduction

1.1 Purpose and Scope

This deliverable reports the final progress of Task 4.1 (Optimised audio capturing through MEMS devices) and Task 4.2 (devAIce platform for audio-visual analysis and voice anonymisation). Both tasks started at M06 and have been completed by M24. An initial version of the deliverable exists under the label D4.1 where the progress of both tasks up to M12 has been reported.

The full development of the MEMS microphones has been completed under T4.1, where the current version includes advanced features compared to the initial one such as supporting up to 8 microphones and edge AI processing. The final results of this new version of the microphones are already available.

Likewise, in T4.2, the new VAD model upgrade has been completed. It has been re-trained according to a novel SOTA architecture and the support of music detection along with speech has been added; this was completed by M12. By M24, the full integration into devAIce SDK was completed, and the shortage observed while benchmarking such as responsiveness and reactivity has been addressed. Furthermore, the latest benchmarking results are reported in this deliverable.

Moreover, the data collection and user annotation applications such as SensMiner and AVER toolkits have been further improved to adapt to the MARVEL use cases, and are now being used or ready to be used. Finally, the steps for building and deploying the audio anonymisation pipeline have been completed, and the description of the workflow is provided in the deliverable.

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 collaboration 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. Moreover, the hardware developed in this task has been distributed and used in different pilot use cases, which will be further explained later in this document.

In addition, T4.1, T4.2, and T3.1 are highly correlated and partners have been working closely. The components developed within these three tasks; MEMS microphones, devAIce VAD, and AudioAnony; will represent the audio anonymisation pipeline, and will be used by the pilots for data acquisition and also to store fully anonymised databases, therefore contributing respectively to WP6 and WP2. Moreover, streaming anonymised audio directly from the edge avoids transmitting speaker-sensitive data and therefore adhering to the privacy requirements and ethics set in WP9. Also, the toolkits developed for data collection and user annotations will contribute to reaching the goals set in WP2.

Finally, the components listed in these tasks have specific KPIs defined in D1.2, the progress of which will be reported in this deliverable.

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 while reducing latency and bandwidth usage.

T4.2 contributes to project objective 1 by achieving KPI-O1-E3-1: *Number of incorporated safety mechanisms (e.g., for privacy, voice anonymisation)* ≥ 3, through building an on-premise audio anonymisation pipeline which will serve as a first pre-processing and data ingestion layer in the MARVEL Big Data pipeline.

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 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 an audiostream sent via Wi-Fi. 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 Operating parameters of MEMS microphones

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 XENSIVTM MEMS microphones, IFAG offers a wide variety of microphones from low cost to high performance. During the course of this project, IFAG has kept researching, testing, and developing new microphones to encompass the necessities of the new applications previously commented.

As explained in D4.1, the pillars of the audio acquisition development chosen at the beginning of the project, was the XENSIVTM IM69D130 MEMS microphone, depicted in Figure 1. This SOTA MEMS microphone presents high standards for power consumption, frequency response, performance, and other acoustic features that can be found in Table 1.

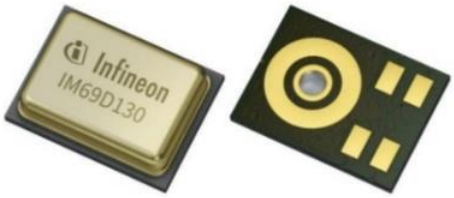

Figure 1: Latest IFAG MEMS microphone IM69D130

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				

In order to supply the growing market, and the requirements to cover the new applications which are more specific, the XENSIV™ family has grown. In the context of this project, two new products are relevant, in particular the new IM72D127 and IM72D128. As shown in Table 2, where the main acoustic characteristics are described, the performance of these two new microphones is lower than the previous version. The main advantage presented by these two new products is the protection against water and dust, having an IP57 certification.

Table 2: Acoustic characteristics IM72D12X

Parameter		Symbol	Values			Unit	Note or Test Condition
			Min.	Typ.	Max.		
Sensitivity			-35	-34	-33	dBFS	1kHz, 94dB SPL, all operating modes
Acoustic overload point		AOP		127		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			67			
	f_{clock}=768kHz			65			20Hz to 8kHz bandwidth, A-Weighted;
Total harmonic distortion	94dBSPL	THD		0.5		%	Measuring 2nd to 5th harmonics; 1kHz. All power modes
	123dBSPL			1			
	127dBSPL			10			
Low frequency cutoff point		$f_{C\,LP}$		40		Hz	-3dB point relative to 1kHz
Group delay	250Hz			113		µs	
	600Hz			23			
	1kHz			9			
	4kHz			3			
Phase response	75Hz			30		°	
	1kHz			2			
	3kHz			-2			
Directivity			Omnidirectional				Pickup pattern
Polarity			Positive pressure increases density of 1's, negative pressure decreases density of 1's in data output.				

As the water and dust protection for the devices developed in this project was planned from the beginning, the novelty introduced by the new microphones developed by IFAG will not bring any extra functionality to the board's development. Therefore, the basis for all the development will still continue to be the IM29D130.

2.1.1 EVAL _FLEXKIT

For quick evaluation purposes, the microphones can be seamlessly connected to an edge device, without directly soldering them, by using the evaluation kit shown in Figure 2.

The EVAL_KIT 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 an I2S to PDM or I2S to PCM converter to reduce the complexity of the interface. It is available for all the three products described above, i.e., the IM29D130, IM72D127 and IM72D128

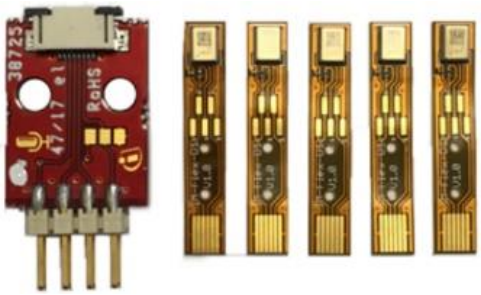

Figure 2: EVAL_IM69D130_FLEXKIT

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 3.

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 GitHub[3].

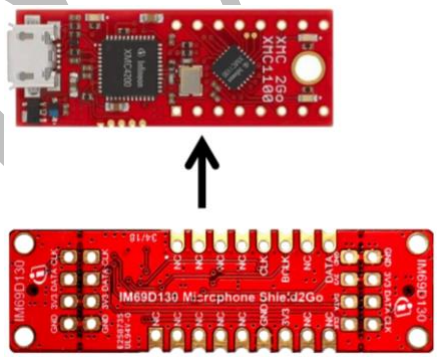

Figure 3: Shield2Go board and XMC 2Go

2.2.2 IFAG Audiohub – Nano

The Audiohub – Nano, depicted in Figure 4, 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

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

- 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

Figure 4: IFAG Audiohub – Nano

The board does all the required processing to transform the PDM input from the two microphones to two channels USB Audio, as depicted in Figure 5. 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.

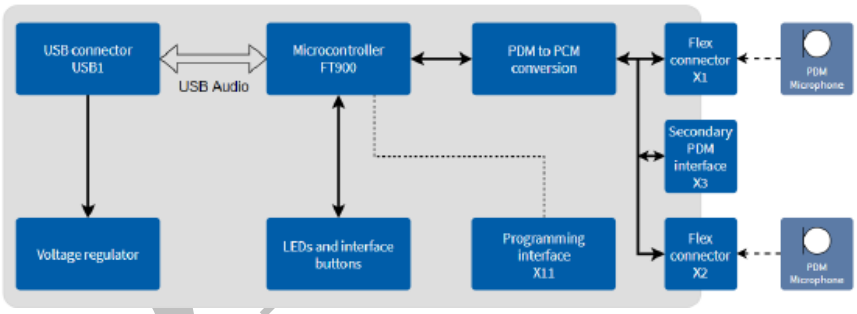

Figure 5: IFAG Audiohub - Nano block diagram

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 6. 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. This board also implements the PDM to USB conversion, based on the different components shown in Figure 7.

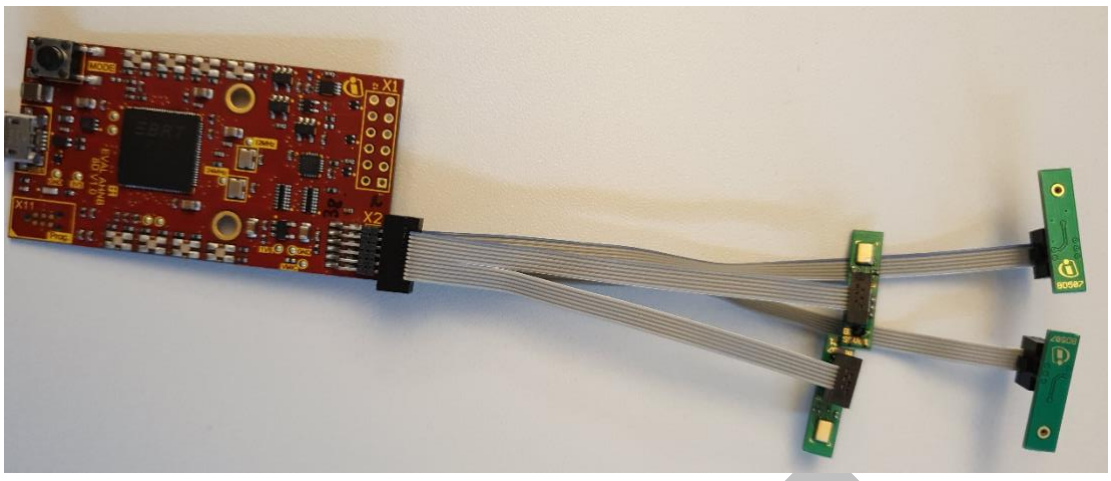

Figure 6: Audiohub - Nano 4 microphones

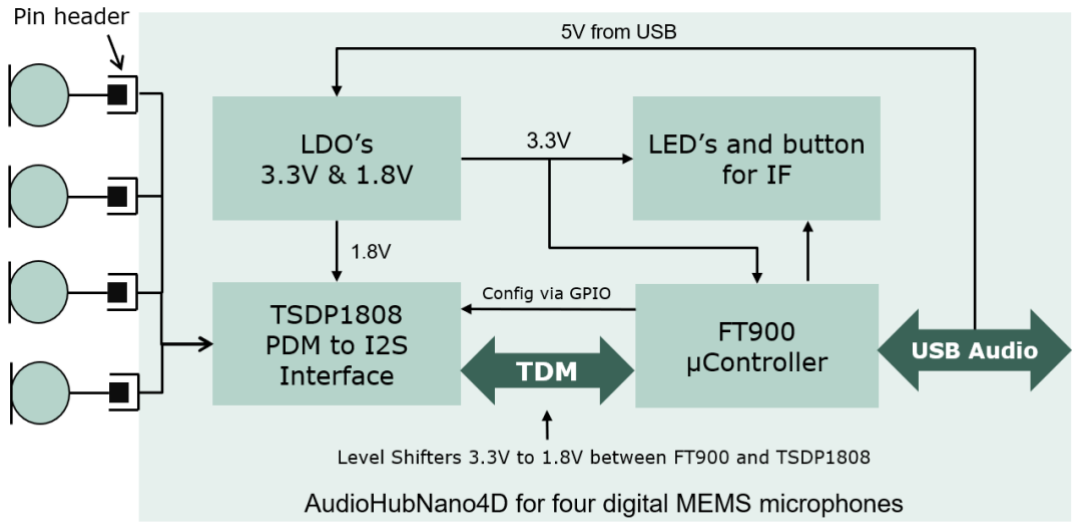

Figure 7: Block diagram Audiohub - Nano 4 microphones

2.2.4 Customised XMOS XK-USB-MIC-UF216

An over-the-counter microphone board and processor[4] was customised for ultra-low latency applications. Instead of the original 7 microphones (Figure 8) 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 (Figure 9).

- 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

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

- Ultra-low latency (~ 5x original speed)

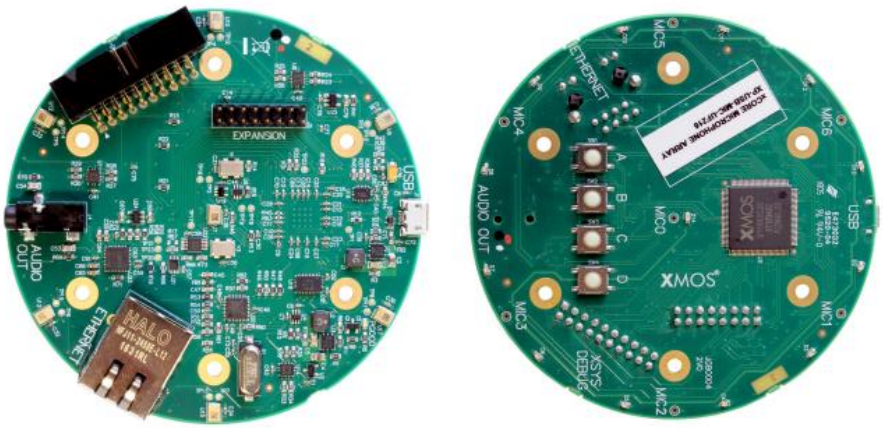

Figure 8: XMOS XK-USB-MIC-UF216 original PCB

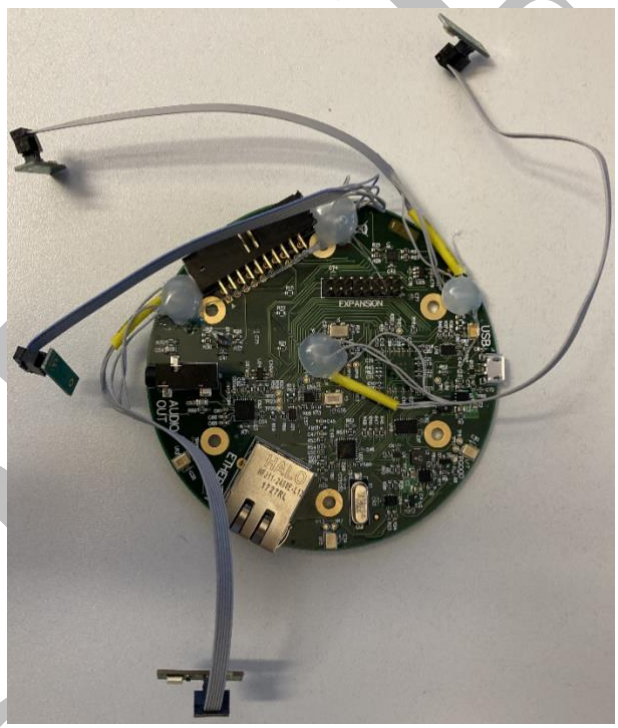

Figure 9: XMOS XK-USB-MIC-UF216 modified PCB

2.2.5 IFAG Audiohub – Nano 8 Mic Version

IFAG has developed an 8-microphone version of the Audiohub – Nano (Figure 10) to meet the requirements needed by the MARVEL project partners. It is a dual-PCB stacked design. It consists of a separated microphone board (Figure 11) and a main board (Figure 11), which is used for data processing. 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
- 8x IM69D130 microphones
- Up to 48 kHz frequency sampling.
- Up to 24bit sampling resolution.
- Audio Data streaming via Wi-Fi
- Wi-Fi Module (Murata 1DX Wi-Fi-Module)
- PSoC 64 processing board

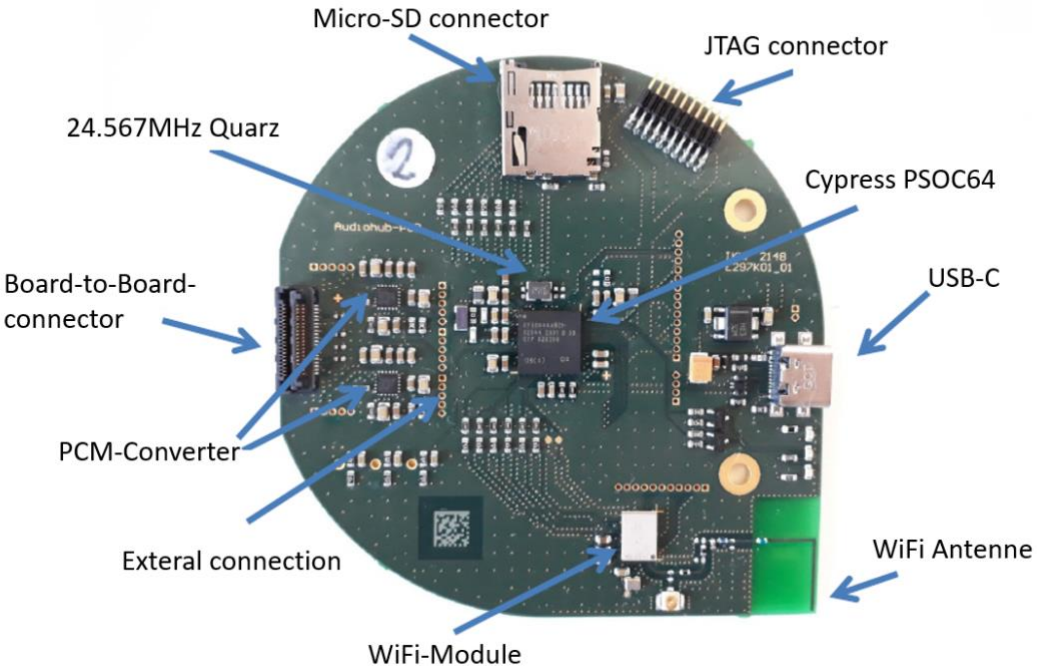

Figure 10: Dual-PCB 8-microphones Audiohub – Nano: Main board

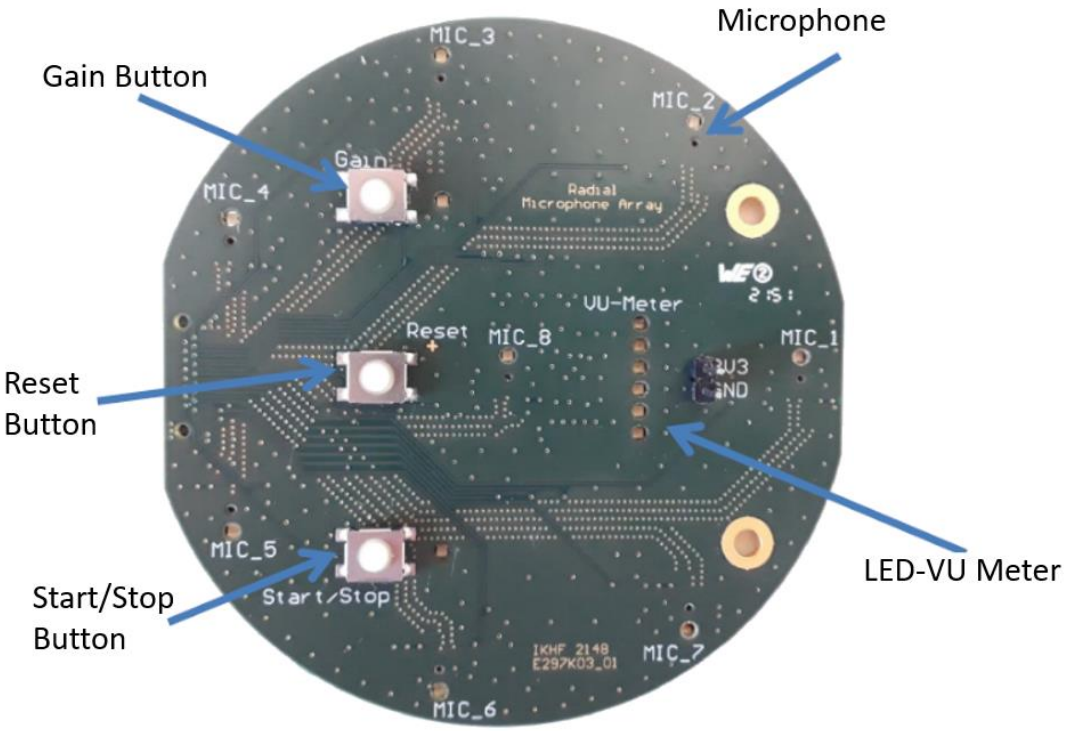

Figure 11: Dual-PCB 8-microphones Audiohub – Nano: Microphone board

The PSoC 64 is the processing unit of the Audiohub – Nano with 8 microphones. Figure 12 shows its capabilities. It will be delivered with a default firmware streaming the raw 8-channel audio data via Wi-Fi to an external cloud. The signal acquired from the 8 microphones can be also stored in an external micro-SD card (i.e., there is no Wi-Fi connection available). The firmware capabilities of this board can be further programmed and debugged for extended functionality via JTAG or SWD.

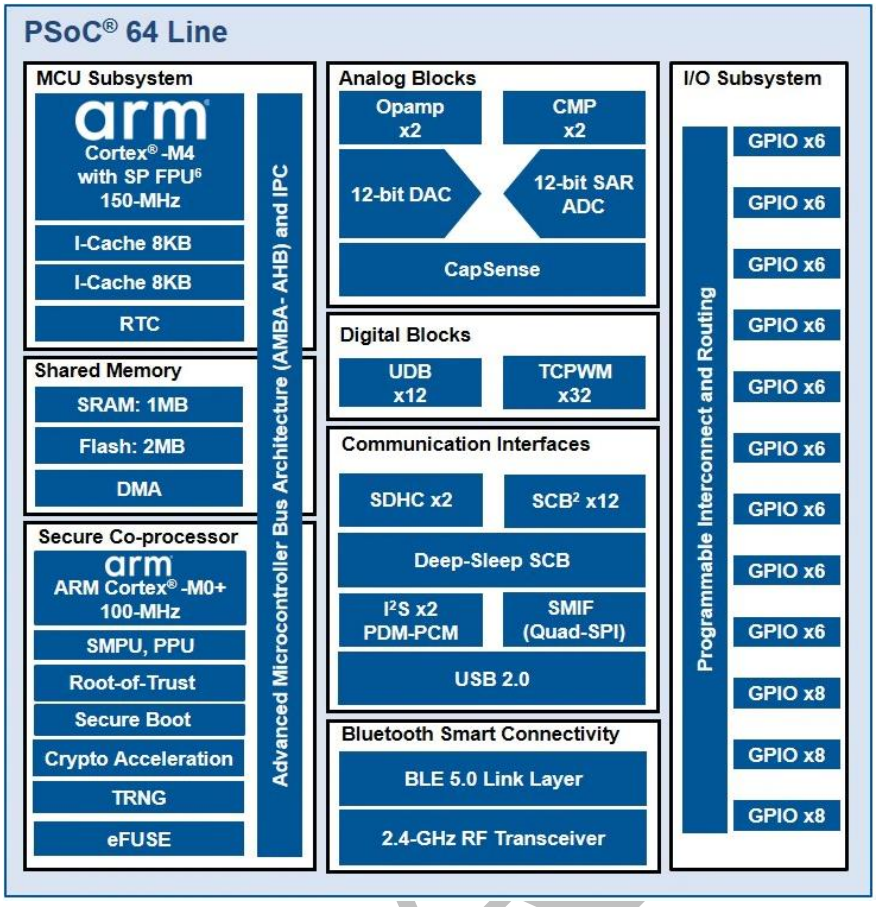

Figure 12: PSoC 64 block diagram

The PSoC family offers the software ModusToolboxTM ML, publicly available online[5], which allows for rapid evaluation and deployment of ML models on all IFAG MCUs, including the PSoC 64 (Figure 13). 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.

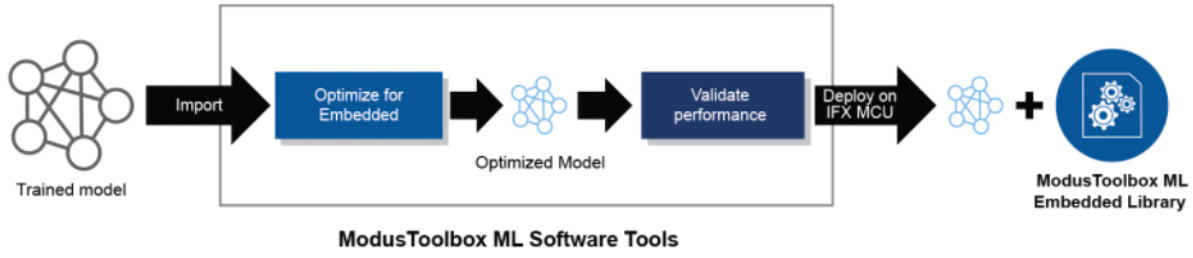

Figure 13: Flow for ML on the PSoC 64 with ModusToolbox

The custom firmware allows the configuration of different acquisition parameters (i.e., sampling frequency or resolution, number of channels, etc.), as well as the Wi-Fi configuration parameters to use in any desired network. This configuration, depicted in Figure 14, can be done via USB serial communication or MQTT.

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

```
loaded settings
WiFi SSID            : MARVEL_AP
Audio channel map    : FF
Audio resolution     : 24
Audio sample rate    : 48000
SD-Card filesize     : 256
WLAN MAC Address : D0:17:69:E8:48:07
WLAN Firmware    : wl0: Jul 18 2021 19:15:39 version 7.45.98.120 (56df937 CY) FWID 01-69db62cf
WLAN CLM         : API: 12.2 Data: 9.10.39 Compiler: 1.29.4 ClmImport: 1.36.3 Creation: 2021-07-18 19:03:20
WHD VERSION      : v2.5.0 : v2.5.0 : GCC 10.3 : 2022-09-23 13:14:02 +0800
```

Figure 14: Audiohub Nano 8mic configuration

Furthermore, a water and dustproof custom housing has been developed for the Audiohub – Nano 8 Mic Version. It will allow the use of this device in pilot scenarios where the placement of the board is exposed to weather's inclemency. Figure 15 shows the housing and the board fitted in its position.

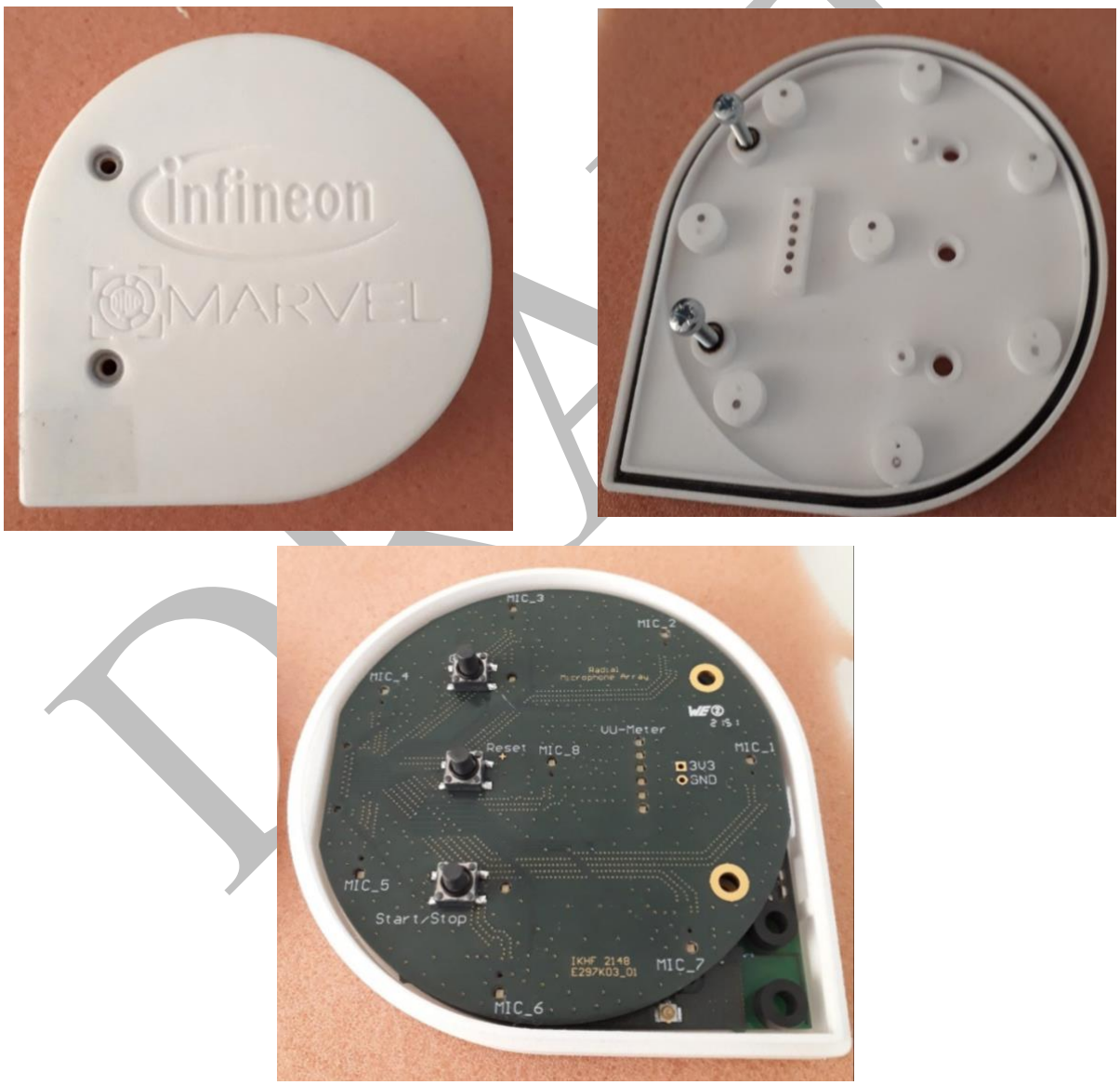

Figure 15: Housing and board assembled

2.3 Pre-processing for Data analytics (basis for light-weight ML models)

The devices presented in the previous subsection set the basis for the audio acquisition and further processing. The definition of a methodology for compressing Audio/Visual models has been one of the main topics of investigation so far. Precisely, the activity focused on defining a methodology for compressing a specific neural network architecture, AVCC, meant for estimating the number of people present in a frame. The AVCC network, designed by AU, takes in input a video frame and audio signal properly pre-processed for extracting the corresponding spectrogram. The overall architecture is composed of three main parts: (i) a VGGish backbone for the audio processing, (ii) the first 13 layers of the VGG Deep network, and (iii) a sequence of fusion layers whose purpose is to combine the audio and visual high-level features to obtain a density map containing the information of the people heads distribution, i.e., a heatmap, to be further counted to get the final response. Given that the layers constituting AVCC are mainly convolutional, the compression method applied to AVCC goes in the direction of identifying and pruning (i.e., set to zero) those that, in probability, are less active during training.

Such a pruning procedure has already been applied successfully to single SOTA Convolutional Networks such as VGG in the pattern recognition task. However, after an intensive sequence of experiments and trials, we observed that although a significant part of the network is removed, the final prediction performance at inference time is unsatisfactory. The quality of the heatmaps suffers from consistent degradation, probably due to the heavy parameter pruning. In light of this, the activity moved on by investigating how to combine a less aggressive parameter pruning with other compatible techniques, like low bit-representation. The main objective is to obtain similar compression performance (around 50% of the original size), limiting the degradation of expressiveness of the final compressed model. Once finalised, such a procedure might be, at least in principle, applicable to the other models defined in MARVEL.

Two of the MARVEL processing components have been re-designed in order to allow processing on low-end microcontrollers. These components could be suitable for deployment on MEMS devices featuring some computational power. The first component is SED@Edge which is capable of performing sound event detection with very limited resources in terms of computation and memory. This is obtained by employing PhiNets networks (Paissan et al [1]) as modular scalable backbones that can be easily tuned using few hyperparameters (Paissan et al [2], Brutti et al [3]). FaceSwap@MCU performs face-swapping on a microcontroller targeting a Kendryte K210, with a RISC V dual core working at 400MHz. It overall consumes less than 300mW, achieving over 15fps with an FID score under 150. The system is based on PhiNets (Paissan et al [1]) applied to GAN. More details about these components are available in D3.1[6] and D3.3[7].

2.4 Deployment in R1 use cases and final stages experimental results

MEMS microphones are currently used in the MT3: *Monitoring of Parking Places* use case, where a public park place is monitored with audio-visual sensors for the detection of specific events. A specific "audio device" has been developed by FBK that consists of 3 hardware components:

- The MEMS microphones connected with the NanoHub board provided by IFAG.

[6] MARVEL D3.1: Multimodal and privacy-aware audio-visual intelligence – initial version, 2022. https://zenodo.org/record/6821318#.Y49yq3bMLIV

[7] MARVEL D3.3: E2F2C Privacy preservation mechanisms, 2022 - [Zenodo link to be released]

- A Raspberry PI board to which the NanoHub is connected via USB and which performs all the audio edge processing.
- An auxiliary board that automatically resets the Raspberry PI when it stops working in case of overheating or software failures. In this way, the component is always reachable via SSH.

Being deployed outdoors (on lamp poles near the cameras), the device is wrapped in a plastic box. The device is connected to the surveillance network of MT for what concerns both internet connectivity and electrical power and it is accessible via SSH once a VPN is active within the MT network. Figure 16 shows a sample of the device currently deployed in MT3 use case and is being currently deployed in the other use cases involved in the R2 prototype.

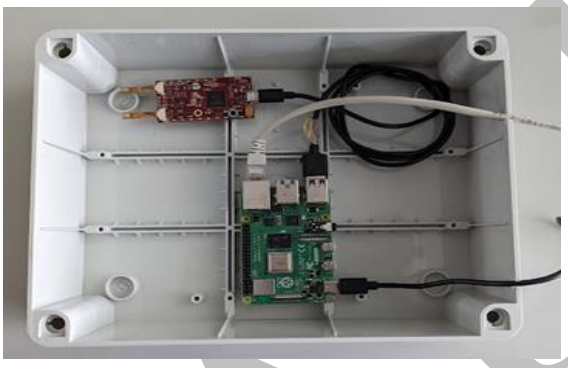

Figure 16: The "audio device" currently deployed in the MT3 use case. MEMS microphones are connected to the NanoHub board by IFAG (red one). The green board is the Raspberry PI. The reset board is not visible here

The multichannel audio stream produced by the NanoHub is read by the Raspberry Pi through the PulseAudio server, at a sampling rate of 16 KHz and 32bit resolution, in mono format. PyAudio is used in Python to handle the audio captured by the microphone. The stream is then processed and streamed using FFmpeg via RTSP. More in detail, the Raspberry Pi hosts the VAD+AudioAnony component that handles:

- the audio stream capturing and re-streaming via RTSP
- the audio anonymisation via AudioAnony
- the voice activity detection via devAIce
- and the related messaging with MQTT

Figure 17 shows the block diagram of the components deployed on the MT edge devices. More details are available in D5.4[8].

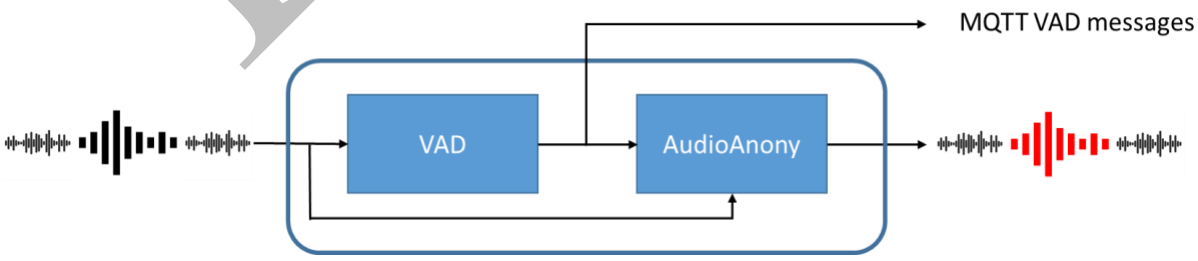

Figure 17: VAD+AudioAnony component deployed on MT edge devices

[8] MARVEL D5.4: MARVEL Integrated Framework – initial version, 2022. Confidential.

In addition, audio signal processing is incorporated as a supporting modality within the UNS drone experiment. While video recording was performed by the camera attached to the drone, audio recordings were collected on the ground using Audiohub Nano development board and MEMS microphones provided by IFAG.

Every Audiohub Nano board was extended by 2 MEMS microphones so stereo recordings were collected. Based on initial experiments and analysis performed after equipment delivery, the board was set to use the normal power mode with the highest gain of 24 dB. The recording process is done using mobile phones since the usage of laptops would be less practical for the recording in public space. Each mobile phone was installed with a recording app which enables stereo recording via USB port (through which development board was connected to the mobile phones). The sampling frequency was set to 44.1 kHz.

The experiment was performed at the Petrovaradin fortress. Three microphones (with the development boards) were positioned approximately 20cm above the ground and were organised in a straight line at the edge of the recording scene. The distance between them was 4 meters. The fourth microphone was positioned in front of the three aforementioned microphones, inside the recording scene. The audio source of predefined audio events (music and anomalies) was positioned in front of the microphones, whereas participants were moving throughout the scene. The setup is shown in Figure 18. The real recording scene is shown in Figure 19.

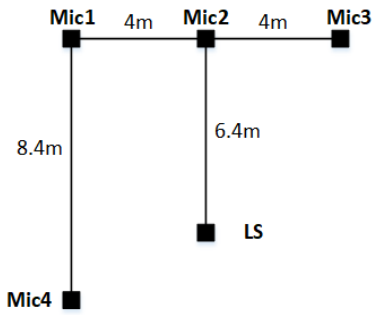

Figure 18: Schematic recording setup

Figure 19: Real recording setup

3 Intelligent audio data analysis and collection

T4.2 aims to provide intelligent audio analysis, and collection and to adhere to the privacy constraints with regard to audio data. These constraints can be summed up in the non-transfer and non-storing of speaker sensitive features, which are all embedded in raw speech. These features can be used to extract valuable speaker-related information. To achieve the necessary privacy measures, the VAD module, interfering in the very first steps of the anonymisation pipeline, has been upgraded and re-developed according to a novel SOTA. The module has also been integrated into the devAIce SDK. This integration will allow the module to function on edge devices, and that way the information will be encoded before being transferred to the next computing layers for either further analysis by the rest of MARVEL AI components or for storing in the MARVEL Data Corpus. This encoding is realised via the intermediate of AudioAnony, the audio anonymisation component. The role of AudioAnony is to anonymise the speech segments detected by the VAD module.

These core modules in the anonymisation pipeline have been fused and deployed on the edge devices attached to the microphones, allowing the speech anonymisation to take place on-premise, ensuring maximum privacy awareness. The anonymisation pipeline also allows, through the VAD module, the detection of overlapping music segments along with speech, allowing for further processing on-edge AI.

In addition, toolkits such as SensMiner or AVER, have been respectively updated and developed to make the data collection and annotation more efficient, allowing the pilots to increase the amount of data, which will later be used either for in-domain training or benchmarking.

3.1 devAIce platform

devAIce is a software development kit wrapping AUD's intelligent audio analytics modules as well as openSMILE, the award-winning toolkit, which can be used for large feature space extraction in more than 6000 dimensions, ready to be fed to ML models. devAIce contains various modules that can be used for different use cases. Such modules are sound event detection, acoustic scene classification, and speech emotion recognition. devAIce is written in C++ and is configured to function on powerful computing nodes as well as high-end edge devices such as Raspberry Pi and Intel Nuc. A compressed version of the toolkit containing just the openSMILE toolkit also exists and can be used on edge devices with limited computational capabilities where only ML models with simple architectures can be deployed.

devAIce exposes an interface in Python, iOS, Android and C. Nevertheless, it can still be used with other programming languages, allowing the user to manually build his own wrapper around the main C interface. This portability and flexibility of deployment make its use straightforward for multiple scenarios.

The SDK, with its VAD module, is a key component in the audio anonymisation pipeline built, together with AudioAnony and MEMS microphones.

3.1.1 A new responsive VAD model

As stated previously, devAIce wraps multiple AI modules, among which is VAD. In order to be suitable for the MARVEL use cases, the model has been upgraded and redeveloped according to a novel SOTA architecture. This architecture is inspired by a research paper published by Lee et al. [4], where a new approach is adopted. This new approach introduces the use of Convolutional Layers as an attention mechanism on top of the Recurrent Neural

Networks (RNN) layers, which is a Long-Short Term Memory (LSTM) layer in our case. Moreover, this attention mechanism focuses not only on the temporal domain like the most commonly used Artificial Neural Network (ANN) based attention modules, but also on the frequential domain, which helps achieving a better modelling of the dependencies between the consecutive frames within an audio signal, by focusing on the parts that contain the most relevant information for Voice Activity, and this on both time and frequency domain. This concept is called Dual Attention. Details on the attention module as well as the overall architecture were reported in the initial version of the deliverable, D4.1[1].

To train the module, a data mixing framework has been set up, which goal is to exploit the currently available internal databases for speech, music, background, and foreground noise as well as other sounds events, by mixing the different audio segments in order to generate a new artificially mixed audio database, simulating the conditions of the audio data that the module will be exposed to when deployed. Furthermore, various augmentations on the audio segments used for mixing have been applied, among which are applying different types of filters (low-pass, high-pass, equaliser, etc.) to simulate different recordings conditions and reducing the gain stage of the audio makes the model suitable for detecting far-field microphone speech.

The mixing framework was developed to be easily scalable, allowing the direct introduction of new data when desired. This means that when annotated in-domain data shall be available, it can be injected into the mixing framework to generate new artificial data and use it for training, this if we find the need to. Figure 20 shows an overview of the data mixing framework.

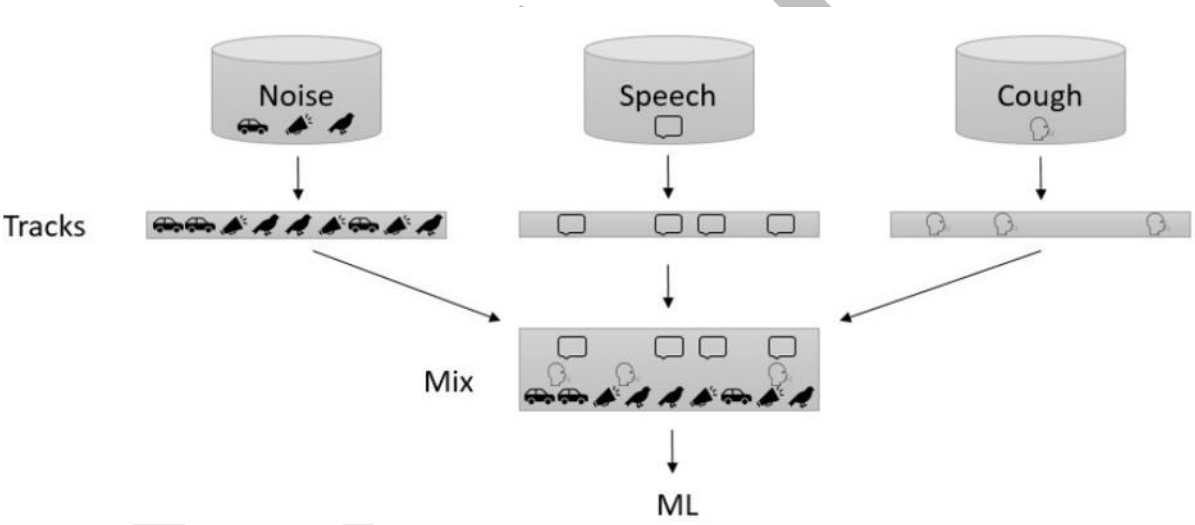

Figure 20: Overview of the data mixing framework

By M12, the first version of the new model, including the new music detection feature, has been delivered to the partners for offline testing, and by M14 the new model's integration in devAIce was completed.

To assess the performance of the model, several benchmarks have taken place, where an increase in performance has been observed, e.g., the UAR scored on the artificially generated database increased from 79.8% to 83.8%. More-in depth benchmarking as well as resource-driven benchmarking were reported in D4.1, where the new model proved to be 4 times faster on multi-core processors.

In the meantime, continuous benchmarking and improvement have been taking place. Among these improvements, we observed increase in the model's reactivity and responsiveness. For context, model's reactivity is described by the ability of the model to adhere to sudden changes

in the context of the audio signal. Such changes include short pauses between syllables in speech, breathing moments in a long speech segment and so on.

The lack of responsiveness in the new model was observed after frame-level benchmarking on VAD Talkshow, a database of various German talk shows. The database contains mainly speech without silence or other events, with a lot of short pauses. These short pauses are often ignored by the new model and classified as speech.

In terms of numbers and scores, such behaviour can be explained by a high percentage of False Positives (FP) rate and a low recall of the non-speech frames. Table 3 below compares the old VAD model (plain LSTM) and the first version of the new dual attention-based model. Bold numbers refer to better performance, where for UAR and Recall, the higher the better, and for FP rate, the lower the better.

Table 3: Comparison of models VAD performance on VAD Talkshow

	UAR	Non-speech Recall	False Positives rate
Plain LSTM Model	**0.89**	**0.80**	**20%**
Dual Attention Model	0.83	0.66	34%

Visually, the lack of responsiveness can be explained by the non-sudden drop of the model scores for voice activity, which is illustrated in Figure 21. The horizontal line in purple represents the ground truth, where the upper bar indicates speech and, the bottom one indicates non-speech. The orange points indicate scores of the new model and in green those of the old model. The scores represent the probability of speech within the audio sequence, and it can be seen that the green scores (old model) tend to be more reactive than the orange ones (new model).

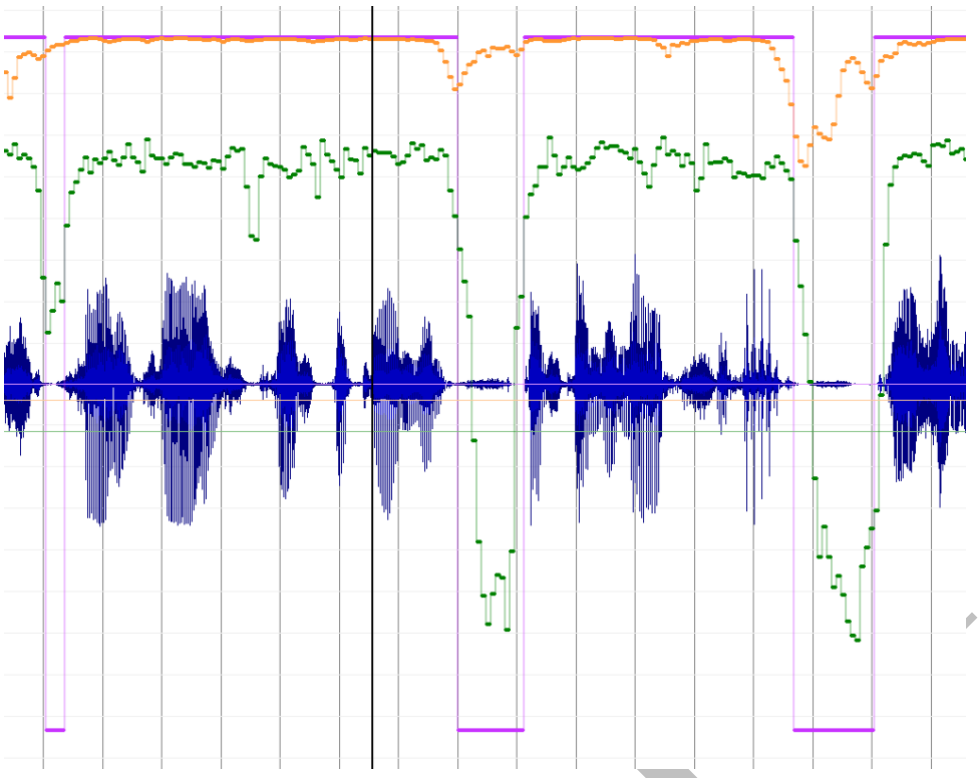

Figure 21: Model's activation scores during a short pause context

To address this behaviour, two approaches have been followed:

- Semi-supervised learning-based approach:

We trust the model to refine the learning base, so we start the supervised learning process, and after a few epochs when the voice recognition begins, we refine the ground truth labels of the training data, where the new refined data will be used for the next epoch. The refinement is based on a simple approach, the model calculates the predictions on the training base frames, and rounds them according to a certain threshold, to either speech or non-speech. As this approach was not successful, theoretical and implementation details will not be discussed further in this deliverable.

- Signal processing-based approach:

Based on the calculation of Root Mean Square (RMS) Energy of the audio signal. This step was adopted as a pre-processing in the mixing framework, where for each speech database used, an energy threshold relative to the maximum RMS value of the sequence is defined, and the RMS sequence per frame is calculated for the whole signal. Then, the frames labelled as speech but whose RMS value is below the defined threshold will be refined and labelled as non-speech.

As stated previously, only the second approach was able to solve the reactivity shortage, for that reason more details will be reported. It is also important to note that a combination of the proposed approaches has been tried, but it ended up yielding worst results rather than simply sticking to the RMS-based approach.

The RMS energy thresholding method is illustrated in Figure 22. The sudden drop in the RMS energy (represented by a yellow curve in the middle grid), is due to a short pause during speech, and the model is expected to learn and behave in the same way. The audio segment illustrated in the figure has already been subject to RMS thresholding to generate the fine-grained speech labels, taking into consideration the short pauses and other similar behaviours that cause a

sudden drop in the RMS value. These fine-grained labels are indicated in the bottom grid of the figure, where a box containing S indicates speech and a blank box indicates absence of speech.

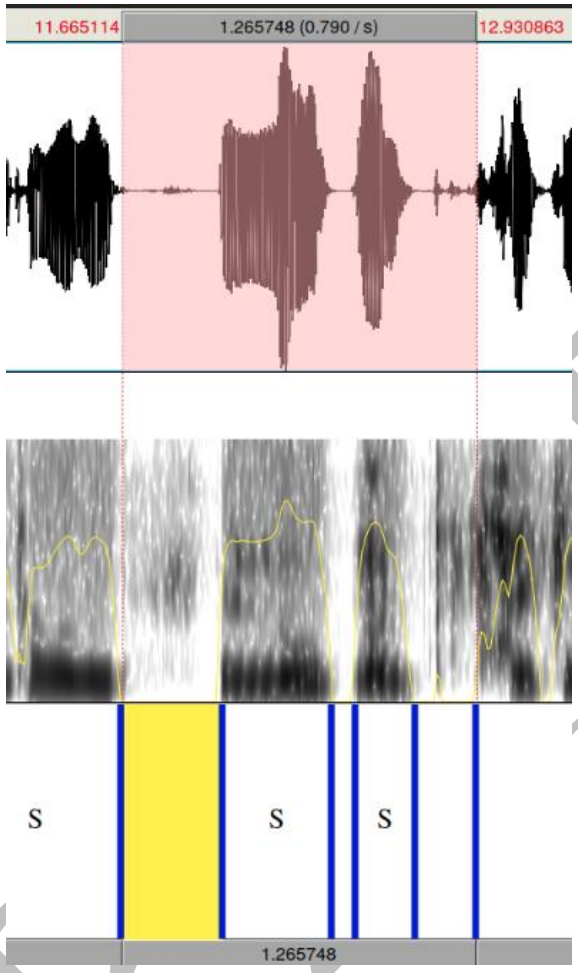

Figure 22: RMS energy and annotations on an audio segment containing short pauses between speech

The difference between the original labels generated before the energy thresholding and those after applying the solution is illustrated in Figure 23, where original labels are represented by the green line and the refined labels by the red curve.

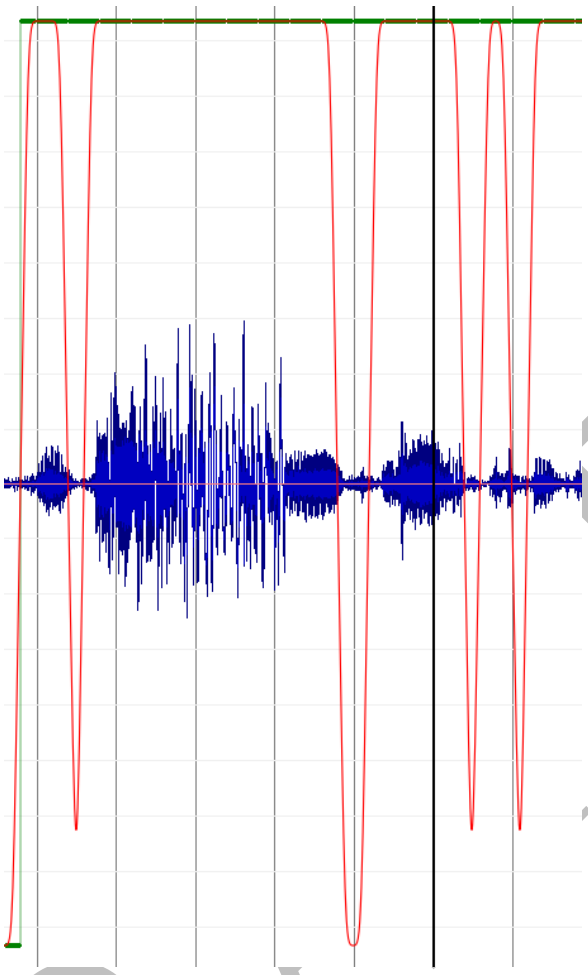

Figure 23: Ground truth labels before and after RMS energy thresholding

This solution, applied to the training data, has allowed decreasing the occurrence of such behaviour and therefore increasing the model reactivity. Numbers and scores describing this improvement will be reported in the next sub-subsection.

3.1.2 Benchmarking in real world settings

To assess the reactivity of the new model candidate as well as the overall performance for both speech and music, a new internal real-world database has been collected. This database was collected through AI SoundLab[9], an AUD's internal application. This tool allows data collection by creating customised surveys from an available template, containing several audio prompts and metadata. Examples of audio prompts in the survey involve asking the user to read from a certain text while playing music in the background and asking the user to play music from different devices. The user is also asked to answer pre-defined questions and provide his opinions about real-world topics. This kind of prompt helps trigger spontaneous speech as an answer. These answers can contain multiple spontaneous sound events like humming, laughing, coughing, and most importantly short pauses taken for thinking or breathing.

This new database has been manually annotated. However, it's been observed that manual annotations can cause bias such as ignoring short pauses and labelling them as speech, plus as the models are operating on frame-level, it is not very trivial to manually determine the exact

[9] AI SoundLab: https://www.audeering.com/products/ai-soundlab

start and end of a segment. As a remedy, the same RMS energy thresholding method that was used to pre-process the training data was also applied in the manual annotations in order to generate fine-grained labels, resulting in manually annotated speech segments being divided into more precise subsegments. Table 4 compares the performance of the old VAD model (plain LSTM), as well as the new model (Dual Attention) before and after RMS energy thresholding on the AI SoundLab data.

Table 4: Comparison of models VAD performance on AI SoundLab data

	F1-score	Recall	Precision
Plain LSTM Model	0.88	**0.93**	0.83
Dual Attention before RMS	0.86	**0.93**	0.81
Dual Attention after RMS	**0.89**	0.92	**0.87**

To highlight the reactivity improvement after the RMS energy solution, Table 5 compares the performance of the different candidates on the VAD Talkshow database using the same metrics as in the previous sub-subsection. Here the capability to detect short pauses increased resulting in the FP rate to drop to 10%.

Table 5: Comparison of models VAD performance on VAD Talkshow after RMS solution

	UAR	Non-speech Recall	False Positives rate
Plain LSTM Model	0.89	0.80	20%
Dual Attention Model	0.83	0.66	34%
Dual Attention after RMS	**0.90**	**0.90**	**10%**

As stated previously, continuous benchmarking has been taking place and two more databases were used, the VAD Toolkit database containing conversational speech between two Korean male speakers recorded in real-world environments (i.e., construction site, bus stop, park and room), and the Artificially Mixed Dataset. Table 6 and Table 7 show respectively the models' performance on the VAD Toolkit and the Artificially mixed dataset, where the new model post-RMS energy thresholding tends to score better than the rest of the candidates.

Table 6: Comparison of models VAD performance on the VAD Toolkit

	F1-score	Recall	Precision
Plain LSTM Model	0.92	0.93	0.92
Dual Attention before RMS	0.90	0.93	0.89
Dual Attention after RMS	**0.93**	**0.95**	**0.93**

Table 7: Comparison of models VAD performance on the Artificially mixed dataset

	F1-score	Recall	Precision
Plain LSTM Model	0.77	0.74	0.82
Dual Attention before RMS	0.84	**0.82**	0.87
Dual Attention after RMS	**0.86**	**0.82**	**0.92**

To benchmark the music detection performance, the AI SoundLab data was used. As the old VAD model doesn't perform music detection, the new model candidate was compared to a baseline model InaSpeechSegmenter [5] operating in "speech, music, and noise" mode. Table 8 shows the comparison, where again the candidate post-RMS thresholding is showing better results overall.

Table 8: Comparison of music detection performance on AI SoundLab data

	F1-score	Recall	Precision
InaSpeech Segmenter	0.81	0.68	0.99
Dual Attention before RMS	0.86	0.78	**1.0**
Dual Attention after RMS	**0.95**	**0.91**	0.99

3.2 Audio-visual data collection

3.2.1 SensMiner Toolkit

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

SensMiner will be exclusively used within the UNS drone experiment and an initial version has been delivered to the pilot in the early months of the project. UNS has conducted indoor laboratory experiments on the SensMiner application and provided feedback for potential improvements.

The application has then been subject to continuous updates based on the continuous feedback received from the UNS pilot, and the major improvements were:

- Added support for the latest android version.
- Added scheduled recordings.
- Added the possibility of listening to the recordings from the history page.
- Changed record directory to the Download folder, allowing a better user experience by avoiding connecting to Android Studio each time to export the data, making recorded audio easily accessible without PC, i.e., directly from the smartphone.
- Fixed settings page.

Figure 24 shows screenshots that illustrate some of the features stated above.

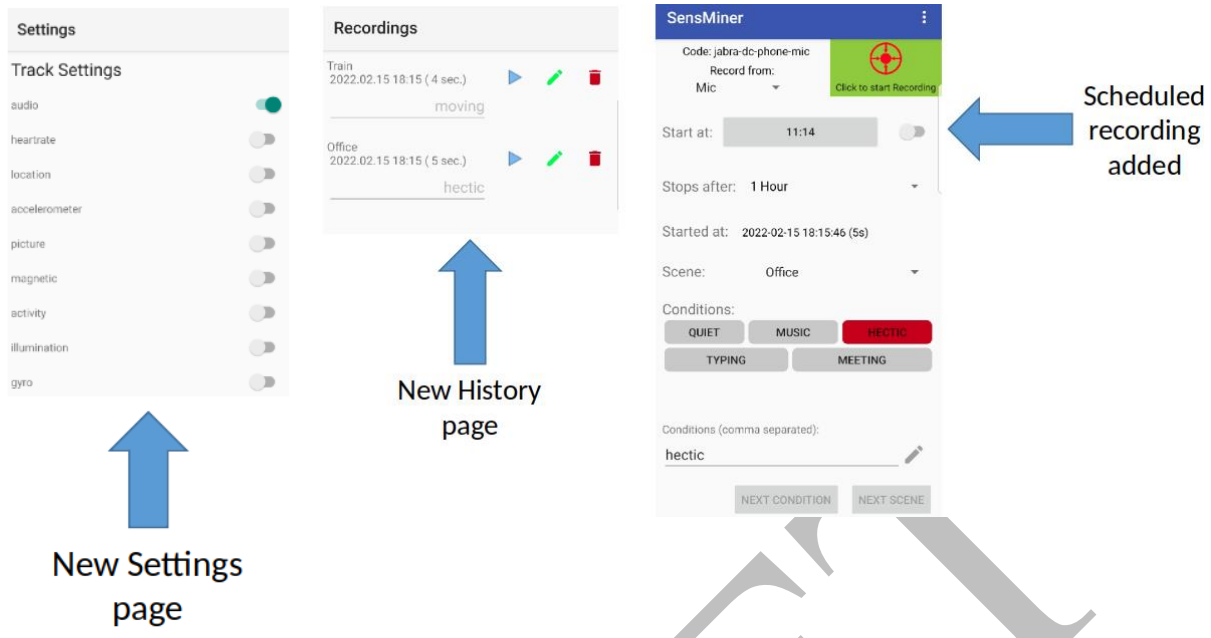

Figure 24: SensMiner major updates

For UNS, it was crucial to have quick and easy access to recorded data in order to check if some staged recorded scenes should be repeated. Having in mind that several MEMS microphones are part of the UNS Drone Experiment setup, supporting various Android operating systems made it comfortable to use our current equipment. These changes made the application suitable for upcoming staged recording phases.

3.2.2 AVER app for data collection

As a part of the UNS emotion recognition use case, the development of the application for the collection of video recordings of emotional expressions is scheduled. This application is built as an extension of a previously developed application for the collection of speech-only emotional expressions in Serbian language [6]. Compared to this version of the application there are two main contributions:

1. bilingual support – Serbian and English
2. video recording support

The application is developed for Android OS based mobile phones and it is accompanied by a dedicated web server, which is used for data storage.

After launching the application for the first time, users are prompted to choose preferred language among English and Serbian. After the language is chosen, privacy policy is shown to the users. Further usage of the application is not allowed until the privacy policy is accepted. After accepting the privacy policy, the basic information related to the users, including their age and gender, is collected (Figure 26a) and sent to the server. During this step, users can also define a nickname. This information would not be used if, for any reason, the user would request to withdraw his/her recording from database.

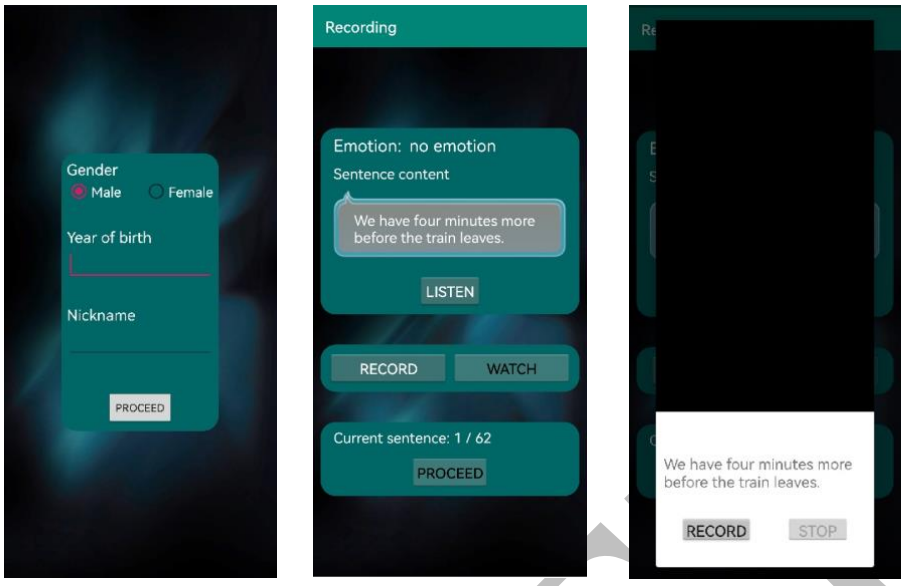

Figure 25: Parts of the AVER GUI: a) user info b) main screen c) recording screen

After basic information is collected, the main screen of the application, shown in Figure 26b, is presented to the users. Privacy policy acceptance and filling basic information will not further be requested from the users on the subsequent application runs.

The main screen contains information about the emotion that is currently being recorded, the orthographic transcription of the current utterance, as well as the ordinal number of the sentence in a given session (one session representing one emotion). Users cannot proceed with recordings of their own rendition of the current pred-defined utterance until they watch the reference recording of the same utterance made by a professional actor. In the same manner, users cannot proceed to the recording of the next utterance until they listen to their own rendition at least once. Such implementation forces one kind of self-assesment of the quality of the recording.

The recording screen, implemented as an overlay to the main screen, is shown in Figure 26c. The central part of this screen is the recoding overview component. Besides the respective buttons for starting and stopping the recording, there is also the orthographic transcription of utterance being recorded.

The recording is performed using the front camera on the device. Since different Android OS devices can support different video resolutions, there is no unique resolution used. The resolution used is determined per each device based on the list of available resolutions in a way that the best available one is used.

The recorded videos are sent to the server in the background using a Wi-Fi connection only (in order to save mobile connection data). If the Wi-Fi connection is not available while recordings are made, the upload process will be retried next time the application is used.

The server side of the system is implemented in C# programming language and connected to the dedicated SQL database, which is used for storing information about users and their recordings.

The examples of screenshots from the application are shown in Figure 27.

Figure 26: Example recordings

4 On-premise audio anonymisation

4.1 Audio Anonymisation with AudioAnony

To maximise the preservation of the citizen privacy using anonymisation, the AudioAnony component must be deployed as close as possible to the microphones. In the MARVEL framework, AudioAnony is deployed on the edge device (i.e., Raspberry Pi) to which microphones are connected via the IFAG NanoHub.

The solution deployed in the R1 prototype is based on McAdams Coefficient [7], a simple approach based on randomised format shifting. The implementation is based on a modification of the code[10] released by the VoicePrivacy Challenge. In order to apply anonymisation only on speech content and preserve the other sound events, AudioAnony is coupled with VAD, which operates on the same device. Details about audio anonymisation, its pairing with VAD, and the plans for the next release, based on the use of neural models for voice conversion with pre-trained large-scale models, are reported in detail in D3.3.

4.2 Deployment and integration of the audio anonymisation pipeline

T4.1, T4.2, and T3.1 end goal is to provide a secure audio ingestion pipeline, which will be the very first layer of the overall MARVEL E2F2C pipeline. This first layer is the audio anonymisation pipeline, composed of VAD, AudioAnony, and MEMS microphones. The workflow will be the following: MEMS will capture the audio stream, which will be first pre-processed by the VAD module to detect the relevant speech and music segments in real-time. Once a speech segment is detected, in parallel, the boundaries of the events will be forwarded to the fog and cloud layers for storing and visualisation, and the speech segment will be processed in a second stage by AudioAnony. Once this second processing step happens, also in real-time, speech information will be anonymised and speaker identity will be safe which ensures that privacy is preserved, only then the audio stream will be forwarded to the next layers for further processing and analysis by the rest of the MARVEL components. Figure 27 below illustrates an overview of the audio anonymisation pipeline.

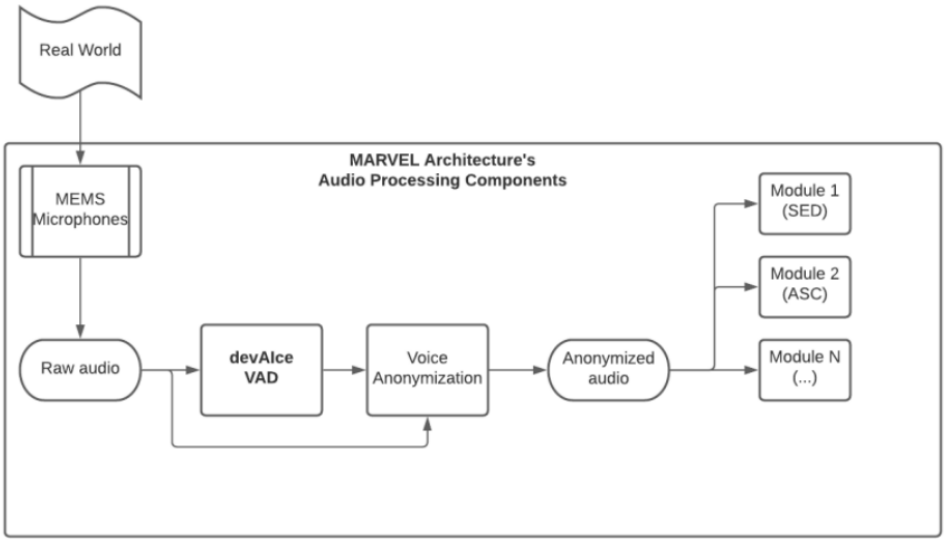

Figure 27: Overview of the MARVEL audio processing pipeline

[10] https://github.com/josepatino/Voice-Privacy-Challenge-2020

To ensure efficient communications between MARVEL's different heterogenous components, as well as security, so no outsider could penetrate through the pipeline, all components were dockerised and deployed as part of a Kubernetes cluster monitored by MARVdash. This way the progress and any occurring vulnerabilities can be easily monitored through a user-friendly dashboard.

Moreover, as an extra privacy measure, the audio anonymisation pipeline was deployed on the edge, precisely inside the chip equipped to the MEMS microphones, which can be either a Raspberry Pi or an Intel NUC processor. This exploits the flexibility and the low resource consumption of devAIce, VAD, and AudioAnony, which were combined into a single component and deployed within the same docker container, where VAD was installed as a python solution and used within the same code as AudioAnony. This approach guarantees better efficiency and lower latency in real-time speech anonymisation by avoiding excessive traffic between both components, such as time establishing and verifying connection between the two different containers, exchanging the information through RTSP and so on. Therefore, the composite components will only have to connect to the RTSP canal to forward the anonymised audio and to the MQTT broker to forward the boundaries of the detected events.

The workflow of the composite component is visualised in Figure 28 below.

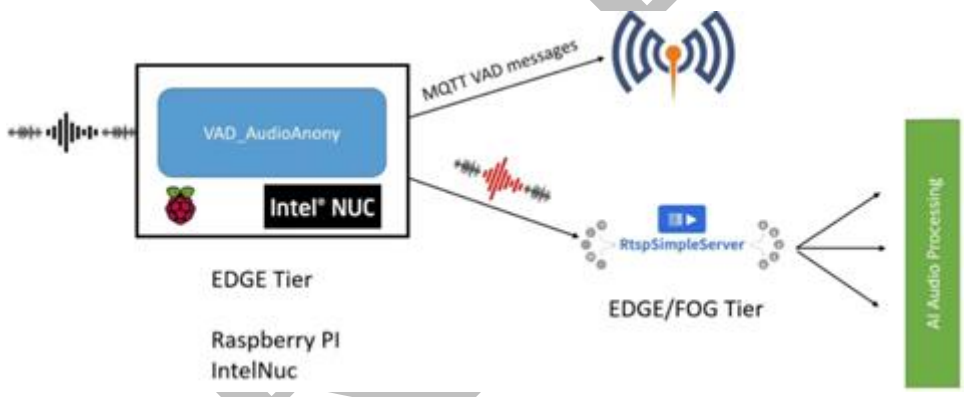

Figure 28: Workflow of the composite component VAD-AudioAnony

To dive into the deployment details, devAIce is made to function on cross-platforms, therefore a different package exists which is made to be deployed on a specific architecture. Different Dockerfiles were created, each for a specific architecture (arm64, armv7 or amd64). Depending on the architecture adopted by the edge device, the adequate dockerfile will be executed. The screenshot in Figure 29 shows a snippet of the dockerfile used to deploy the container on an arm64 architecture.

```
# Start FROM Ubuntu image https://hub.docker.com/_/ubuntu
#FROM --platform=$TARGETPLATFORM alpine:3.15 AS build
FROM --platform=$TARGETPLATFORM python:3.10-slim-bullseye AS build
# Install linux packages
RUN apt update && apt install -y gcc portaudio19-dev python3-dev screen nano ffmpeg libsndfile1
# Install python dependencies
COPY requirements.txt .
RUN python3 -m pip install --upgrade pip
RUN python3 -m pip install pyaudio
RUN python3 -m pip install --no-cache-dir -r requirements.txt
# Create working directory
RUN mkdir -p /app
WORKDIR /app
# Copy contents
COPY FondazioneBrunoKessler-devAIce-SDK-3.4.0-2022-02-23 /app/FondazioneBrunoKessler-devAIce-SDK-3.4.0-2022-02-23
COPY Audioanony /app/Audioanony
RUN python3 -m pip install /app/FondazioneBrunoKessler-devAIce-SDK-3.4.0-2022-02-23/bin/python/devaice-3.4.0-py3-none-linux_aarch64.whl
RUN chmod +x /app/Audioanony/launch
RUN chmod +x /app/Audioanony/rtsp-simple-server-arm
```

Figure 29: Code snippet of the dockerfile used to deploy VAD-AudioAnony on arm64 architecture

The docker containers created through the docker file need to be part of the Kubernetes cluster in order to be able to communicate with the rest of MARVEL architecture and forward the anonymised audio streams as well as the model inference result. In some use cases, such as the UNS pilot, it was possible to expose the container network and make it part of the cluster. This was done through the creation of a Kubernetes configuration file labelled "manifest file", which is a YAML file containing the relevant instruction and metadata for the deployment of a container inside a cluster. Figure 30 shows a snippet of the YAML file created for deployment on the UNS edge devices.

```
# rtsp server yaml
apiVersion: v1
kind: Service
metadata:
  name: $NAME
spec:
  type: ClusterIP
  ports:
  - port: 8554
    targetPort: rtsp
    protocol: UDP
    name: rtsp

  selector:
    app: $NAME
---
apiVersion: apps/v1
kind: Deployment
metadata:
  name: $NAME
spec:
  replicas: 1
  selector:
    matchLabels:
      app: $NAME
  template:
    metadata:
      labels:
        app: $NAME
    spec:
      affinity:
        nodeAffinity:
         requiredDuringSchedulingIgnoredDuringExecution:
          nodeSelectorTerms:
          - matchExpressions:
            - key: Layer
              operator: In
              values:
              - UNSEDGE1
```

Figure 30: Code snippet of the Kubernetes configuration file used to deploy VAD-AudioAnony on UNS Edge 1

In the use case of the MT pilot, the edge device (Raspberry Pi) equipped to the lamp pole, wasn't allowing an external connection through its network, and therefore it wasn't possible to make the audio anonymisation pipeline deployed within the Raspberry Pi as part of the Kubernetes cluster. As a workaround, as the edge device was accepting network connection only from FBK workstation, the forwarded anonymised audio as well as the MQTT messages

were forwarded to the rest of the MARVEL architecture with an extra step, through forecasting via FBK workstation. This way, the anonymised audio will reach first FBK workstation, which will forward it instantly to the rest of the MARVEL components, allowing to keep the security measures imposed via the MT network.

The outcome of this on-edge deployment, allowed to create a joint innovation resulting from the collaboration of T4.1, T4.2, and T3.1, namely "*on-premise lightweight embedded device for real-time audio streams anonymisation"*. This joint innovation has been accepted by the EU Innovation Radar[11].

The composite component was subject to several integration tests to ensure the right synergy and interaction between the audio anonymisation pipeline and the rest of the MARVEL components. These tests were carried out continuously, bugs have been observed and by M24, they have been addressed resulting in a successful deployment and interaction with the rest of the overall pipeline.

One of these components is SmartViz, designed and developed by ZELUS. It is an extensible software with a wide application scope ranging from Data Forensics Analysis to Big Data analytics. The functionalities of SmartViz include advanced visualisations of detected events, demonstration of the related statistical data via a variety of visualisation widgets, data filtering capabilities, and flexible interfaces. It provides a multipurpose platform for data visualisation and transformation that constitutes the Decision-Making Toolkit (DMT) for the MARVEL project. The DMT is the interface between the end users and the MARVEL framework. At this stage of the project, the scope of the DMT includes advanced visualisation of information that serves in a total of five use cases in the smart cities of Malta, Trento in Italy, and the University of Novi Sad in Serbia. The DMT is able to consume a variety of data coming through different workflows inside the MARVEL infrastructure and visualise the outputs of the included components.

The VAD component is visualised in two use cases, MT3 and UNS1. In MT3, the useful outputs of VAD are represented in two different visualisation widgets, in the Temporal representation (Figure 31) and in the Details widgets (Figure 32). The functionalities of the DMT's visualisation widgets are described in detail in D4.3[12]. In the Temporal representation widget, the depiction of the speech events that are detected by the VAD component across time assists the user in understanding the timeline of their occurrence in a monitored area, whilst the Details widget allows the user to drill into all the available information.

[11] https://www.innoradar.eu

[12] MARVEL D4.3: MARVEL's decision-making toolkit – initial version, 2022. https://zenodo.org/record/6821280#.Y2pUXezP0UE

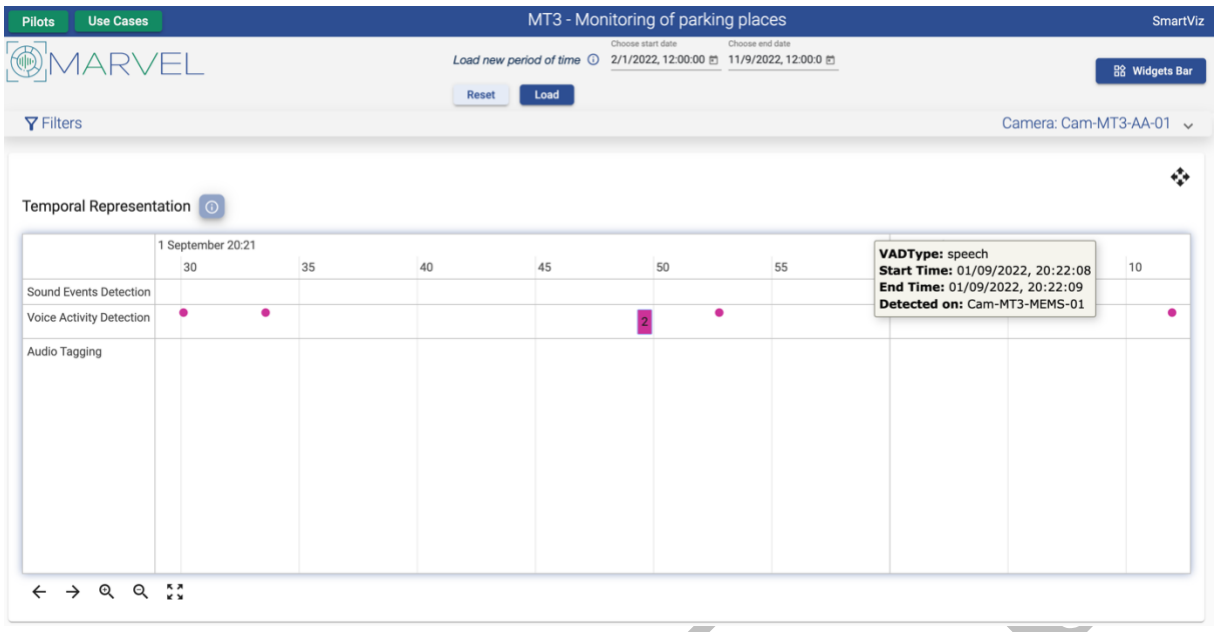

Figure 31: VAD Representation in Timeline widget in SmartViz for MT3

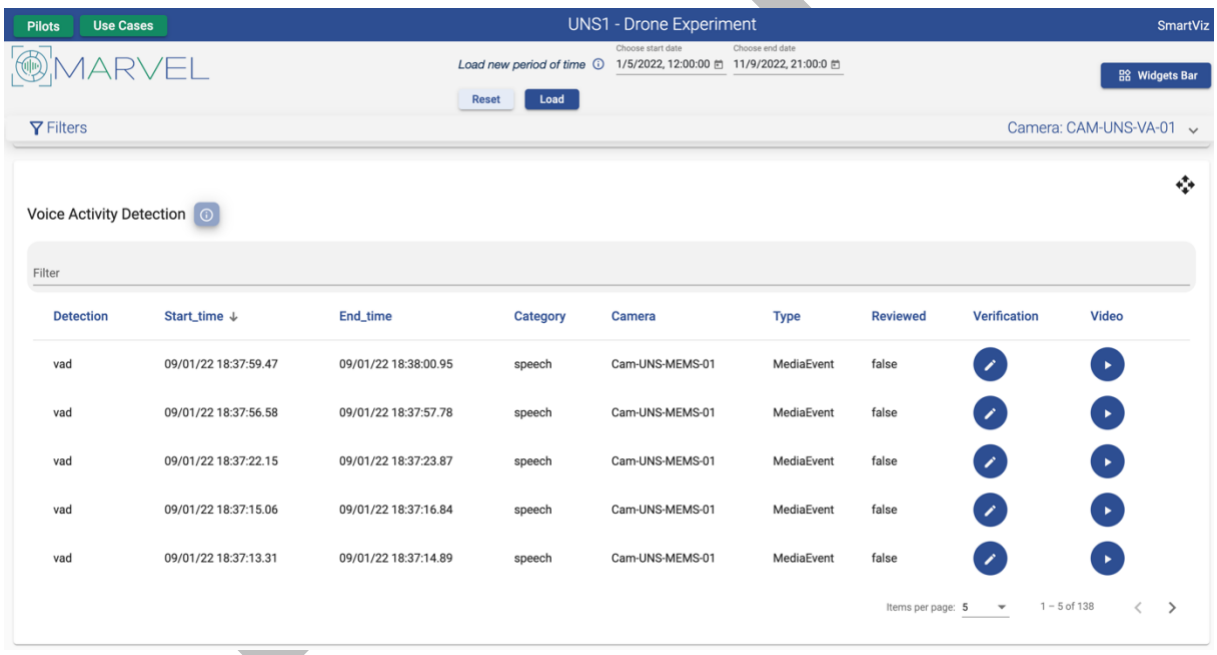

Figure 32: VAD Representation in Details widget in SmartViz for UNS1

5 KPIs

5.1 Project-related KPIS

- KPI-O1-E3-1: Number of incorporated safety mechanisms (e.g., for privacy, voice anonymisation) ≥ 3

  There are three different aspects where safety and privacy are considered in the MARVEL framework. These aspects are Audio and Video data privacy and Network safety, and for each aspect, one or more safety mechanisms are already being exploited, which leads to the overall number of tools used surpassing the defined threshold of 3 safety mechanisms, resulting in the successful achievement of the KPI.

  For Audio anonymisation, the strategy was to first detect the sensitive information in the audio stream, which refers to the speech segments within an audio stream. This is done by the VAD model which was recently updated and re-developed under T4.1 to function in the specific use cases acoustic environments. Once these segments are identified, the baseline was to simply remove them from the stream, which results in privacy preservation but with a loss of information. However, with the recent developments done in T3.1, the speech segments are instead subject to special processing resulting in the conversion of voice and therefore the anonymisation of the sensitive speech features such as speaker identity. Furthermore, both processing techniques (VAD and anonymisation) were optimised to take place on edge devices, ensuring the transfer of only anonymised and non-sensitive audio data to the next layers of the project pipeline (fog and cloud).

  For Video anonymisation, the strategy involves the use of VideoAnony. The applied processing by VideoAnony in the early stages of the project was the detection of the faces and car plates within the video stream, followed by their anonymisation via the intermediate of blurring techniques. With the recent development of the GAN-based AI models, face blurring has been replaced with face swapping, which this led to preserving valuable information such as facial expressions. Moreover, various efforts have been conducted to reduce the resource consumption of the models, this was done through model quantisation, and also by developing a light-weight solution suitable to be deployed on edge devices. Further details on both Audio and Video anonymisation are reported in D3.3[7].

  Finally, for Network safety, the end-to-end framework security is achieved through EdgeSec VPN while the device safety is ensured by EdgSec TEE. Details of the development of these two components can be found in D4.2[13].

- KPI-O1-E1-2: Increase of data throughput and decrease of access latency by 10%

  Currently the number of microphones has been increased from 2 to 8, and the custom firmware developed allows the live transmission of all the audio channels wireless to the next node. It increases the amount of data sent while reducing the latency.

- KPI-O1-E3-3: Video and voice anonymisation expected to improve by at least 10%

  As the original description of the KPI does not indicate what the 10% improvement refers to, a reasonable interpretation has been instead conducted to define the meaning

[13] MARVEL D4.4: Security assurance and acceleration in E2F2C framework – initial version - https://zenodo.org/record/6821254#.Y6r-23bMLIV

of the metric. This led to the KPI being divided into two different aspects, a 10% computational complexity reduction towards edge deployment for video anonymisation; and a 10% increase of the amount of information in the audio content compared to complete speech removal. For both aspects, the KPI has been successfully achieved.

For Video anonymisation, the 10% computational complexity reduction towards edge deployment was achieved via the quantisation of the large AI models (GAN-based architectures) and by developing a light-weight real-time solution suitable for deployment on low-end microcontrollers.

For Audio anonymisation, the 10% increase in the amount of information about the acoustic scene was achieved thanks to the use of voice conversion. Instead of complete removal of the detected speech segments, the segments are subject to voice conversion which results in keeping the acoustic information intact while altering the speaker identity.

Table 9: Project-related KPIs concerning the audio anonymisation pipeline

KPI ID	KPI Description	Strategy	Related Task	Related Component
KPI-O1-E3-1	Number of incorporated safety mechanisms (e.g., for privacy, voice anonymisation) ≥ 3.	**Audio:** 1) Detect voice segments using VAD 2) Filter out sensitive audio content with AudioAnony 3) End2end on-edge processing **Video:** VideoAnony as a safety measure **Network:** 1) End-to-end framework security with EdgeSec VPN 2) Device security with EdgeSec TEE	**T3.1** **T4.2** **T4.3**	**devAIce VAD** **AudioAnony** **VideoAnony** **EdgeSec VPN** **EdgeSec TEE**
KPI-O1-E1-2	Increase of data throughput and decrease of access latency by 10%	Development of the edge device featuring MEMS microphone with edge processing capabilities. Development of algorithms to decrease the latency based on increasing the capabilities of the processing directly at the edge.	**T4.1** **T4.2**	**MEMS**
KPI-O1-E1-3	Video and voice anonymisation expected to improve by at least 10%.	**Video:** Quantisation of video anonymisation AI models A compressed model suitable for deployment on microcontrollers **Audio:** Voice conversion-based AI models	**T3.1** **T3.5** **T4.2**	**AudioAnony** **VideoAnony**

5.2 Asset-related KPIs

- MEMS-Microphones

  The MEMS-Microphones component has been improved by the development of new devices with higher capabilities, both in processing resources and number of microphones. The impact of this development will be observed mostly during the RP2 when the new 8 microphones board are planned to be deployed in different pilot scenarios.

- SensMiner Toolkit

  The SensMiner application has been subject to several tests and hands-on experiences, but it has not been used yet within the concrete use case yet. It will be exploited in the upcoming year in the UNS drone experiment. The first exploitation is planned for Q1 2023 if the weather conditions allow it, otherwise in Q2 2023. The results shall be reported in D6.3, to be released in M36.

- devAIce VAD

  devAIce VAD was benchmarked on multiple datasets and overall, a 10% improvement on the chosen metrics (F1, precision and recall) was achieved. An example is the artificial mixed dataset where the F1 score increased from 0.77 to 0.86 achieving an 11% improvement, precision from 0.82 to 0.92, achieving a 12% improvement and recall from 0.74 to 0.82, scoring an 11% improvement.

  With regards to resource consumption, the RAM usage is minor for both old and new models (<20MB), however, for CPU usage, the inference speed on an input of 1449s long, on a multi-core CPU, decreased from 5.44s with the old model to 1.37s with the new model, reducing the CPU time by 75%.

Table 10 shows the details of the KPIs set for each component.

Table 10: Component-related KPIs concerning T4.1 and T4.2

Component	KPI	Metric	Expected result	Achieved result
MEMS Microphones	Performance	Distortion Sensitivity Phase tolerances Frequency roll-off Multichannel synchronisation	High performance	The final version of the component is equipped with 8 microphones achieving higher performances
	Robustness	Usability on drones, intersections, and public spaces	Can be implemented in all MARVEL use cases	The component can be used in all use cases.
SensMiner Toolkit	Amount of collected data	Data amount in Giga Bytes	Audio data increased by 10%	Component not yet exploited.
devAIce VAD	Voice detection	F1	10% improvement	11%
	Computational complexity	RAM and CPU usage	10% improvement	75% CPU time/usage

	Robustness	Precision, recall	10% improvement	12% Precision – 11% Recall

6 Conclusion

This deliverable reports the full progress of Task 4.1 and Task 4.2 up to M24. Several methods and components have been either improved or newly developed, in order to achieve optimal audio-visual capturing, analysis, and voice anonymisation.

In the context of 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 or Wi-Fi. The development of new acquisition devices with more channels (up to 8 synchronised audio channels) and features (Edge AI capabilities) has been accomplished and the new devices have been distributed to the partners for evaluation and fulfilment of the MARVEL use cases.

With regards to Task 4.2 which covers various aspects; intelligent audio analysis, collection, and voice anonymisation, progress has been achieved on different aspects and goals were reached. Starting with devAIce‘s VAD module, which was updated and re-trained according to a novel state-of-the-art method, extended with a music detection feature and was subject to continuous improvement to address the different shortages observed such as reactivity and responsiveness. The model was also integrated into devAIce, allowing the capability of edge processing making it possible to combine it with AudioAnony and MEMS to build the audio anonymisation pipeline, resulting in a novel joint innovation; on-premise audio anonymisation pipeline. This pipeline was deployed on edge devices for MT and UNS pilots and integration tests with the rest of the MARVEL components were carried out successfully. In addition to that, SensMiner has been subject to continuous updates, and the most recent version was delivered. AVER application development was also carried out, and both toolkits will be used in the upcoming year for data collection and annotation.

7 References

[1] Paissan, F., Ancilotto, A. and Farella, E. *"PhiNets: a Scalable Backbone for Low-power AI at the Edge"*. ACM Transactions on Embedded Computing Systems 2022.

[2] Paissan, F., Ancilotto, A., Brutti, A. and Farella, E. *"Scalable Neural Architectures for End-to-end Environmental Sound Classification*". ICASSP 2022

[3] Brutti, A., Paissan, F., Ancilotto, A. and Farella E. *"Optimizing PhiNet architectures for the detection of urban sounds on low-end devices"*. EUSIPCO 2022.

[4] Lee, J., Jung, Y. and Kim, H., 2020. "*Dual Attention in Time and Frequency Domain for Voice Activity Detection"*. INTERSPEECH 2020, (pp. 3670-3674).

[5] Doukhan, D., Carrive, J., Vallet, F., Larcher, A., and Meignier, S. "*An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality"*. ICASSP 2018.

[6] Suzić, S., Nosek, T., Sečujski, M., Popović, B., Krstanović, L., Vujović, M., Simić, N., Janev, M., Jakovljević, N. and Delić, V. "*SEAC: Serbian Emotional Amateur Cellphone Speech Corpus"*, 2022.

[7] Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., Evans, N. "*Speaker Anonymisation Using the McAdams Coefficient"*. Interspeech 2021, (pp. 1099-1103).