# Deaf people name recognition

## Hardware Architectures for Embedded and Edge AI

SinghMonti Team - Dilpreet Singh e Mauro Monti

POLITECNICO
MILANO 1863

# The core idea

Our idea starts from a need of anyone: knowing when someone is calling us independently from the situation or place in which we are.

In our case we focused on deaf people. In particular on their need to be able recognize when someone is pronouncing their name in a familiar situation or not.

Our solution consists in a system that allows to perform the keyword spotting with the name of the deaf user. More over, the system is able to recognize the voice of a specific person, in order to provide to the user a further information regarding who’s pronouncing the specific name.

# Development of the project


For our project development we relied on the use of Edge Impulse. We used it to collect data, train and test our model and, in the end deploy it by means of an Arduino library. From the library we decided to start from an available example, the “microphone_continous”, which has been modified in order to reach our purpose.

# Our Dataset

For the composition of our dataset we relied on the Speech Commands Dataset* from Tensorflow and Google that already made available a consistent number of 1s samples for the “Marvin” keyword.

We manually registered samples through Edge Impulse in order to add the component of “marvin” samples relative to the specific person registered as a relative of the user.

We introduced a series of samples for “silence” situation. More over we have selected (randomly) a bunch of words different from the main keyword, from the others available from the Speech Command dataset*.

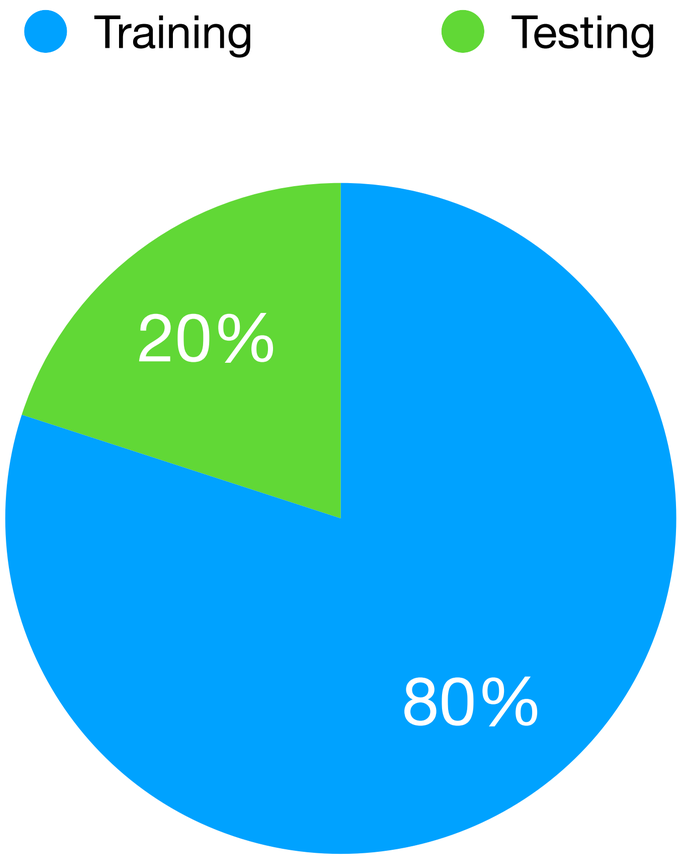
*Speech command dataset developed by TensorFlow and AIY teams @ Google

# Dataset organization

**Labels :**

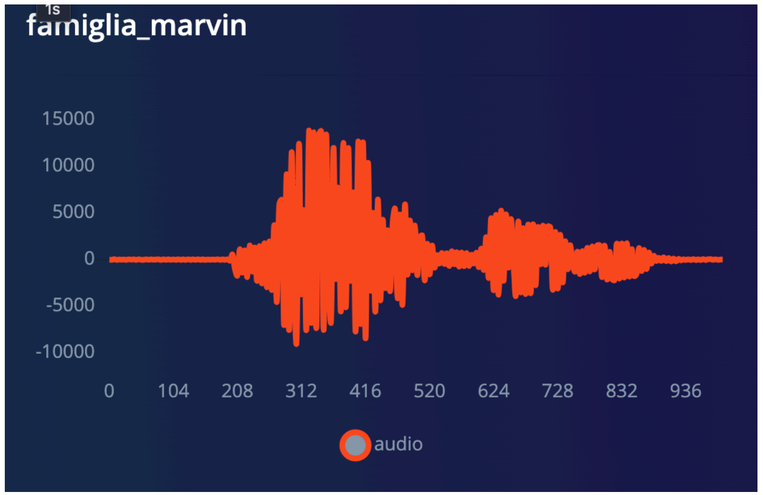
**famiglia_marvin ( 288 )**
**marvin ( 1393 )**
**silence ( 1579 )**
**unknown ( 1389 )**

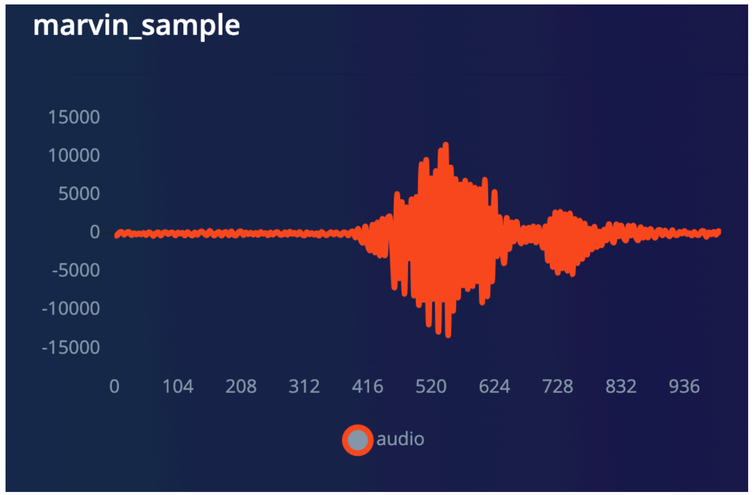
These ones represent the labels of our dataset with the relative quantity of available samples.

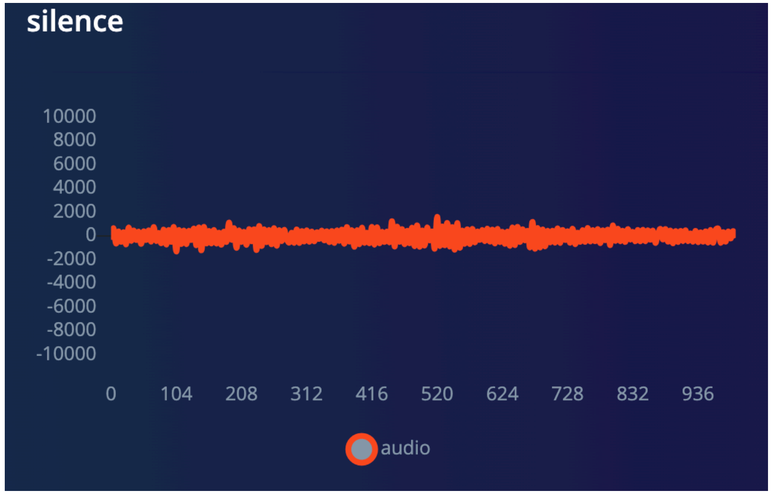


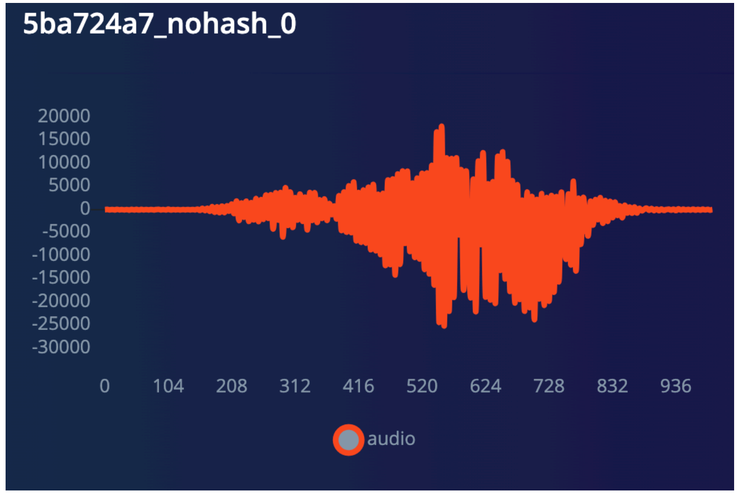

Organization of datas

# Dataset organization



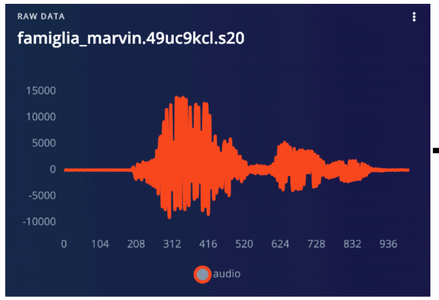




POLITECNICO
MILANO 1863

# Impulse

## Chain for data feature extraction

Raw Data




Time series data with a window size of 1000ms And sampled a 16K Hz

**Processing Block**
**MFCC**

Extracts features from audio signals using Mel Frequency Cepstral Coefficients, great for human voice.

**Classifier**

Learns patterns from data, and can apply these to new data. Great for categorizing movement or recognizing audio

POLITECNICO MILANO 1863

# Impulse

## Processing block: Mel-frequency cepstral coefficients(MFCC)

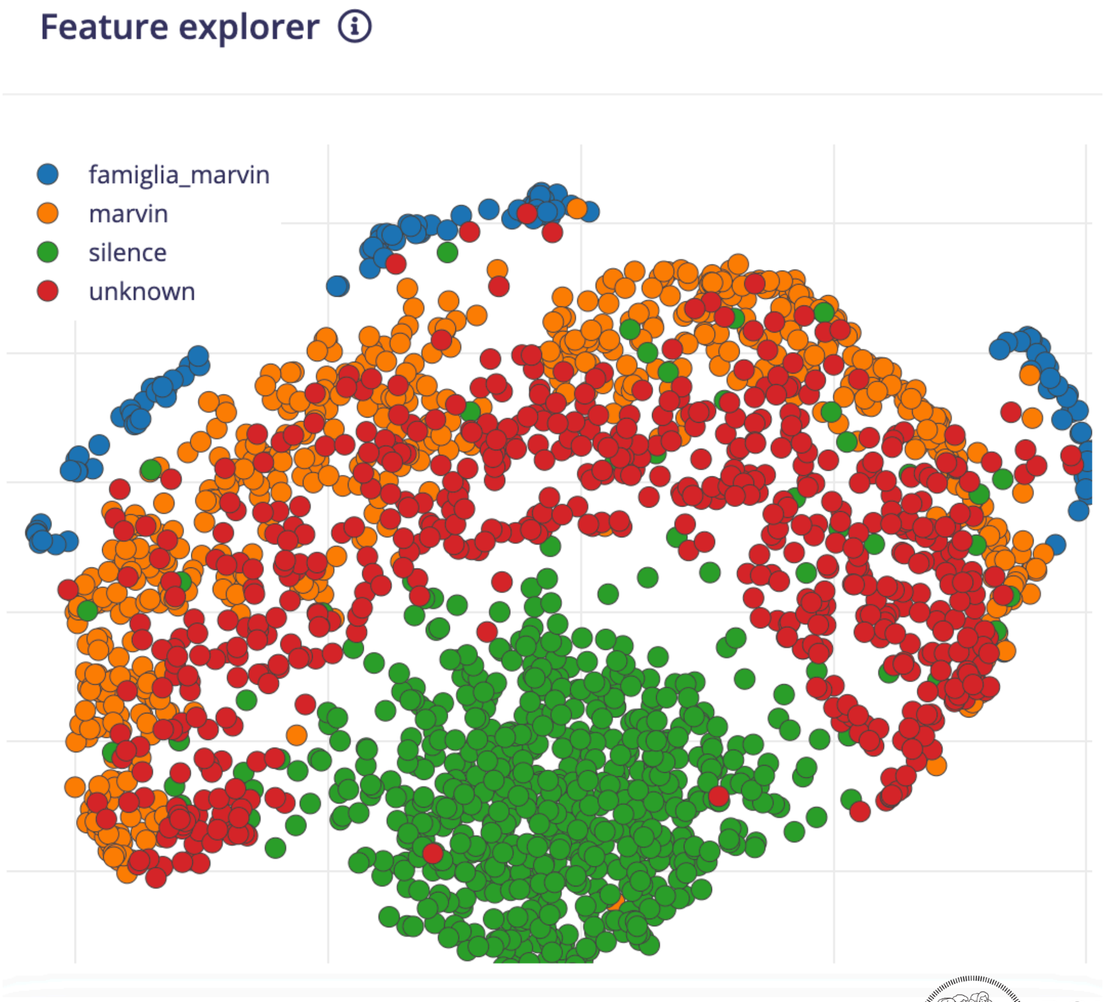
MFCC:
widely used technique for extracting the features from the audio signal.
MFCC processing block takes an audio snippet, **reshapes the information to be more like how our ears hear**, and then extracts key features (coefficients) that describe the unique characteristics of that sound.

# Impulse

## Feature extraction results

We can see how the data separates into different clusters.

We notice how the silence cluster is well separated from the other samples and how there is a bit of overlapping between the unknown and the “marvin” samples.



# Impulse

## Classifier

The classifier has the architecture in figure

Targeted for our architecture:

Arduino Nano 33 BLE Sense

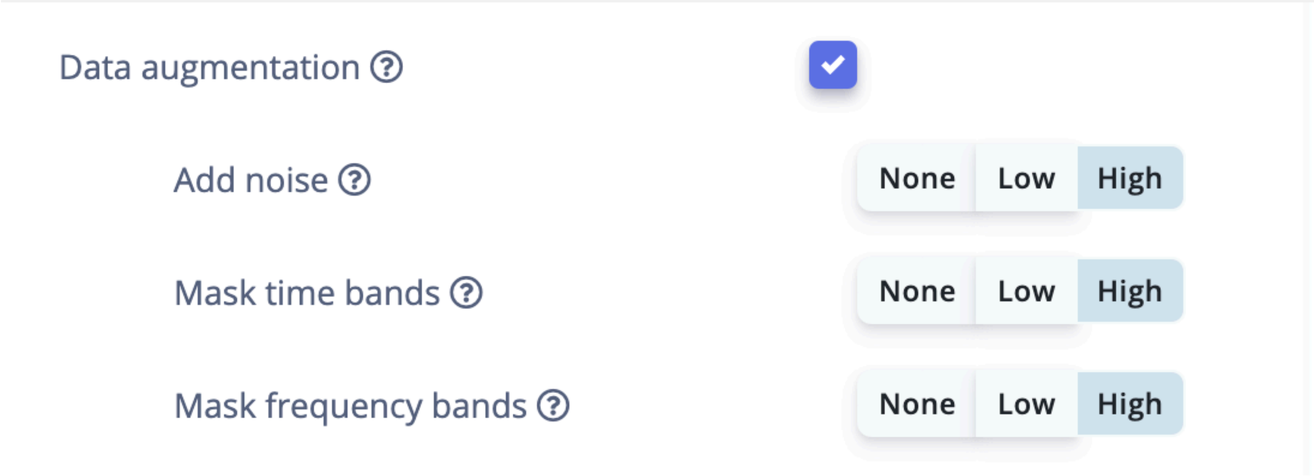

POLITECNICO MILANO 1863

# Impulse

## Training

100 training epochs with a learning rate of 0.005

Data augmentation in order to add noise and mask time and frequency bands.




POLITECNICO MILANO 1863

# Classifier results

ACCURACY
95.6%

LOSS
0,14

**Confusion matrix** (validation set)

| | FAMIGLIA_MARV | MARVIN | SILENCE | UNKNOWN |
|---|---|---|---|---|
| FAMIGLIA_MAR | 92.3% | 7.7% | 0% | 0% |
| MARVIN | 0.3% | 91.5% | 0% | 8.2% |
| SILENCE | 0% | 0% | 99.7% | 0.3% |
| UNKNOWN | 0% | 2.8% | 1.0% | 96.2% |
| F1 SCORE | 0.95 | 0.94 | 0.99 | 0.94 |

Float Classifier Results

ACCURACY
95.6%

LOSS
0,14

**Confusion matrix** (validation set)

| | FAMIGLIA_MARV | MARVIN | SILENCE | UNKNOWN |
|---|---|---|---|---|
| FAMIGLIA_MAR | 94.2% | 5.8% | 0% | 0% |
| MARVIN | 0.3% | 91.5% | 0% | 8.2% |
| SILENCE | 0% | 0% | 99.7% | 0.3% |
| UNKNOWN | 0% | 3.1% | 1.0% | 95.8% |
| F1 SCORE | 0.96 | 0.94 | 0.99 | 0.94 |

Quantized (Int8) Classifier Results

POLITECNICO
MILANO 1863

# Quantization

**“Edge Optimized Neural”**

Employment of EON tool

Int8 bit quantization

POLITECNICO
MILANO 1863

# Deployment

We exploited the microphone_continous example offered by Arduino's library imported from Edge_Impulse. In this way we have been able to establish a serial communication in order to retrieve the results of the inferences of our model running on Arduino Nano 3 3BLE Sense.

Once we tested the model we introduced a series of threshold ruled conditions for which the board turns on the different led lights.

We imposed that for an accuracy equal or larger than 70% for **marvin_famiglia** label, the board should trigger the red led in order to notify the user that the registered person has pronounced "marvin" keyword.
In case an unknown person pronounces the keyword, the green led is triggered.
In case of unknown word, the blue led is triggered.

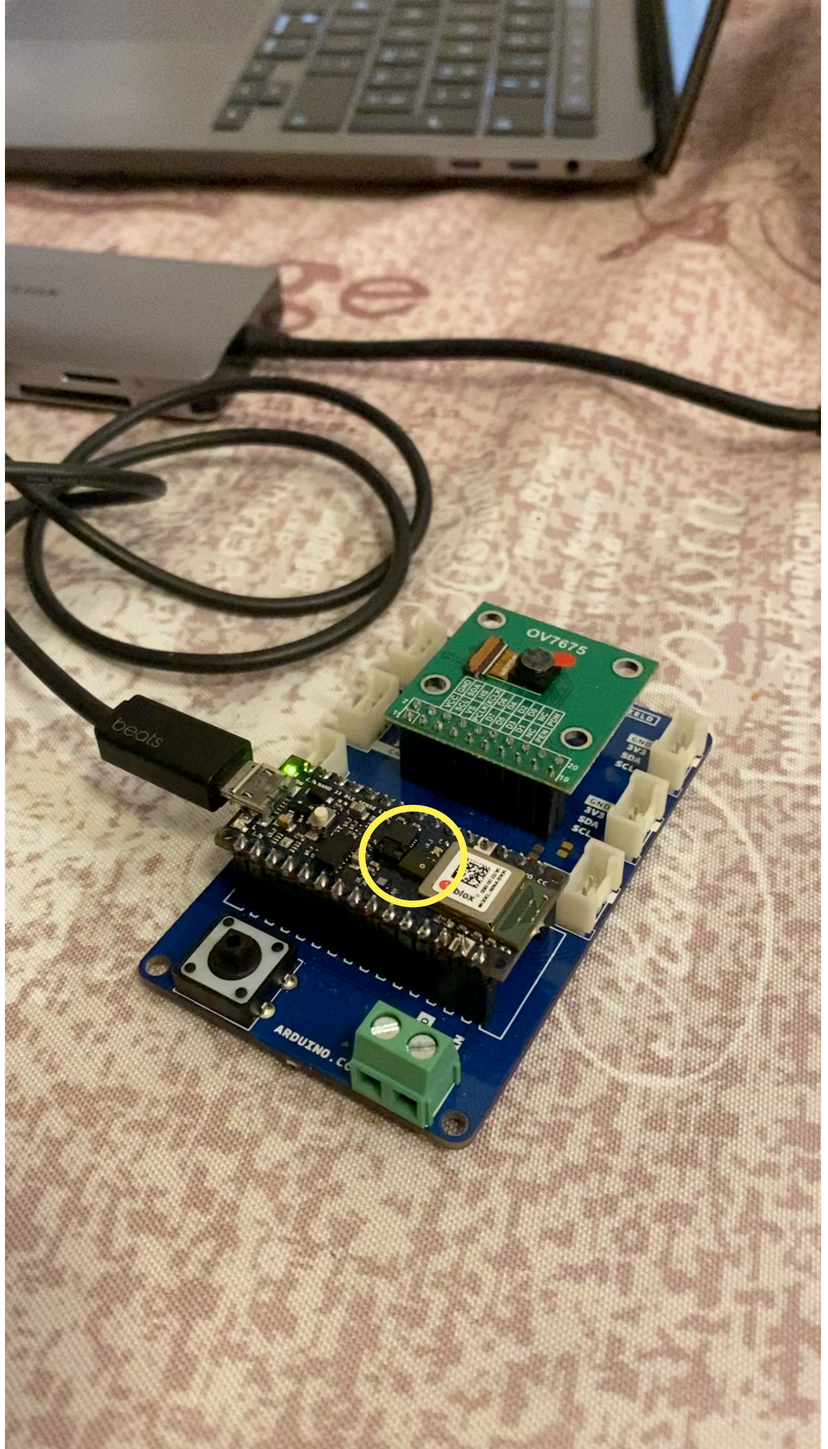
Code provided on GitHub repository

# Live Demo

**Blue** when “unknown”

**Red** when “marvin_famiglia”

**Green** when “marvin”




Video available in GitHub repository

POLITECNICO
MILANO 1863

# Project Ethics

Such system allows people with deafness diseases to achieve the possibility to be acknowledged about people calling them by name.
There could be situations in which the user is not able to directly see the person who’s trying to attract his/her attention, for example a danger one.
Moreover, the aim of this project is to give back to the user part of his independence, being able to perceive what or who is surrounding him in different places or occasions.
The advantage of such project is the possibility to work in offline conditions, offering much more privacy for what concerns users and his/her relatives datas.

# Project Market

In the world almost 360 million people are subject to deafness disease. In Italy almost 900 thousand people are in such condition. The purpose of such system is to provide the final user a specific and fundamental skill for everyday life.

This kind of system has been thought to be embedded in a wearable object such as a watch or smart-clothes. The way in which the system is supposed to provide the feedback (in this case implemented by means of a led) is through vibration patterns. (For example, in a family of 5 people where 1 person is subject to deafness, 4 users can be registered as relatives. 1 vibration is user_1, 2 vibrations is user_2 …)

Although such system may be very user specific and, therefore, it should need many customized audio samples in order to be trained for the final customer, it represents a not replaceable tool in order to improve deaf person life. In fact, many systems for people with disabilities need to be specifically customized for the use by a single person. Even if this represents a difficulty in the mass sell of such a product, it still represents a primary need solving tool.

# Thank you for your attention.

POLITECNICO
MILANO 1863