Eötvös Loránd University
Faculty of Informatics

Improving house security using an IoT infrastructure

Kitlei Róbert
Prof. at ELTE
Marchese Maurizio
Prof. at Trento
Szilágyi László
Customer Manager at Ericsson

Mirko Morandi
Computer Science

Budapest, 2016.

Abstract

The recent increase of power of computation combined to the decrease of power consumption of devices led to an explosion of smart devices for all the possible purposes, and it's called the *Internet of Things* (IoT).

The IoT includes all devices that has an interface to the real world and are connected to the Internet, such as: *Smart Fridges, Smart TVs, Thermostats, Alarms, Cameras and etc.* These devices were born to simplify normal people's life, let's take the smart fridge as example: it will alert you when the food you have in the fridge is going to expire, the thermostat will try to save money applying smart strategies for heat consumption. One of the most important application of IoT to people's life is *House Security*: exploiting the IoT concept to improve the house security with smart sensors like cameras, leap motion sensors and microphones. House Security has to be of concern, because house invasion is an always trending crime. However these systems are not always perfect and the most common problem are false positives, for the which there is an active branch of research focused on.

The emergence of devices connected introduced another problem: heterogeneity between IoT devices. As the popularity of smart devices increased more and more companies started producing their own different ecosystems. Communication between different ecosystems is another research topic which will analyzed in this document.

Acknowledgement

...

Contents

Chapter 1

Introduction

Internet of Things(IoT) based surveillance systems has been proved to be effective during the time, catching the burglar committing the crime and helping the authorities to catch him[1]. Nonetheless these systems are not yet perfect, and during this document we will investigate on some optimizations with respect to the *integration* with other ecosystems and the *reduction* of false positives.
Here follows the main structure of the thesis:

Chapter two introduces the technologies used in our scenario, with a brief introduction on the techniques adopted to face the various problems.

Chapter three illustrates the main architecture proposed during the document and its implementation.

Chapter four shows the validation of the architecture proposed in the former chapter, showing the real benefits of the approach.

1.1 Motivation

In this document we try to illustrate the common problems that arises when introducing a new IoT system in an existing environment. Let's define it as an average person house, we'll define the scenario details later. The first motivating problem taken in consideration is the *integration* of a new system in a house where there's already a different coexisting ecosystem. The idea behind our research in this field is to improve the existing solutions with an innovative and different approach. Nowadays there are many standalone solutions for IoT surveillance, which acts separately from all the other components in the house. We address this isolation, trying to create a system capable of interconnecting with other existing components in the house. There are currently 4 billion of connected devices [2] as of 2016, and the forecasts says they will be 13.5 billion in 2020. This means a growth in the heterogeneity of companies and products, which makes a must the interoperability between devices from different producers.

Most of the common Smart devices are built to work specifically with their own application, without any external support which leads to a loose coupling between devices. This problem arises when introducing new *systems* in a house where there are is already a different ecosystem. The inability of a user to switch or adopt a new technology is also called *vendor lock-in*, a common commercial practice to retain users. However research is moving forward this direction investigating in new innovative and efficient methods to solve this problem, and there are already some approaches which will be discussed in the next chapter. *False positives* is another key issue in our investigation on the improvement of a house monitoring scenario. The reason behind this choice lies in the connection with the "idea" of a house with an alarm connection to a third party security agency. We need to have a real alarm when something is happening, reducing the probability of false alarms and not wasting time calling the police or alerting some vigilance agent.

1.2 The Scenario

In our document we will often refer to an imaginary scenario of a typical house of an average person. It is important to denote this detail since most of our solutions are designed for such a peculiar situation, and they may not be as efficient or innovative in a different scenario.

The main idea is to have a surveillance system in a typical *American style* house, with different floors and rooms, each of them equipped with different and various smart devices. Having a clear idea of the scenario helps to focus more on the identification of issues while designing the systems rather than facing and solving problems at the last moment due to some unpredictable variables of a too varied scenario.

1.3 Calvin

Calvin is an open-source and cloud oriented framework for IoT applications. Calvin is built to scale well, offering the opportunity to run it on both small devices and more powerful machines in order to use the full power of cloud computing. It's distributed perspective follows the principle of *run it where it's most needed*,exploiting the high interoperability with hardware of small devices combined to the more powerful execution of server machines. Calvin aims to be easy to develop, using an actor model architecture where actors are simple but reactive entities interacting with each other.

1.4 Microservices and IoT

The Microservice architecture is an innovative modeling pattern that aims to solve a well defined class of problems: scaling. The idea behind microservices is to split the architecture on different machines usually communicating through a RESTful interface. In microservices the functionalities are exposed through an API interface to the external world using the *gateway pattern*. Microservices share many key aspects with the *microkernel* architecture pattern, offering a better reliability, resiliency, portability, extensibility, flexibility and easier and faster environment for development and testing. With the advent of the need of interoperability, *microservices* seems to hold the key for this problem. Vendors have published their protocols, and have exposed API's to their various hubs. A MicroService can serve as an adapter between various protocols. It can be lightweight and disposable, both desirable traits in a rapidly evolving environment.[3]

1.4.1 M2M: Machine to Machine

Most of common smart sensors, in order to be *smart* they need to provide a form of connectivity: let it be BLE (Blue-tooth Low Energy), Wi-Fi or RFID. *M2M* is trending with the raising of IoT, requiring a higher number of devices to be interconnected without any human interaction. Different devices means different protocols, which introduces difficulties in the communications between each other. The typical approach is to define a set of translators which are able to deal with both sides, also called *Gateways*. Gateways are high-level objects that knows the various protocols needed to communicate and provide these knowledge as a service to whom has the necessity to access the functionalities provided by the device.

In our architecture each gateway is a device exposing a service using a *RESTful* architecture, for a higher simplicity of use.

1.4.2 RESTful interfaces

The communication between different devices is a common and dated problem in *Computer Science*. Many approaches has been introduced during time, all of them with their relative pro and cons:

- **Remote Procedure Calls** used to execute procedures on a remote machine, with the advent of *CORBA* it was also possible to do that on machine with a different technology stack on it.

- The **Message Oriented Middleware** used queues to exchange messages to different platforms using different technologies. This subsequently led to the development of the common *ESB* technology.
- **SOAP** a high-level protocol for communication between web services over many popular application layer protocols such as HTTP or SMTP.

All of them are good candidates for the construction of a distributed system, but it's better to adopt a single protocol for communication instead of many to avoid over complicating the problem. We choose to use the *RESTful* architecture, which better suits our scenario. Our decision complies with the current trend in the implementation of microservices, where REST is the typical choice for exposing functionalities. Its simplicity made it very popular and widely adopted, making REST one of the "standards" for communication between services on the web, mainly due to its capability to work over *HTTP* without any additional infrastructure.

1.5 False positives

One of the most common problems that occurs when an alarm is armed are false positives. In most of the cases it can be just a pet triggering the alarm, some other times may be just yourself when you forget to deactivate the alarm.
Although this problem may look relatively unimportant, it assumes importance if the alarmed is linked to a surveillance company which is supposed to look after our house. That's why during this thesis we will take an effort to address this common issue with different approaches.
The false positive detection is calculated as a probability of an event happening, influenced by many factors. Each factor is one of the approaches adopted below, which will have a positive or negative influence on this factor, determining if with a certain precision a detection can be classified as a *false positive.*

1.5.1 Approaches

Here's a list of the proposed approaches taken in consideration during this document

Face detection Detecting a face when the alarm is triggered is a useful way of discriminating between a false positive and a real intrusion. However this approach has some limitations given by the position of the camera and the speed of the shutter which impacts the image quality. Its importance differs if used in combination with other different *false positives* reduction systems.

Face recognition Face recognition is just an improvement of the above technique, which would allow us to recognize a familiar face in the system and reduce significantly the probability of a false alarm. This approach however is affected by the same problems from the former one.

Speaker recognition This approach involves machine learning to learn and recognize familiar voices during an intrusion. Knowing with a certain precision that the voice belongs to someone familiar will drastically drop the probability of a possible false detection. Furthermore this approach does not suffer from the same problems as the former two approaches.

Dictation recognition People's biometric signature also includes the way they do talk. It is actually possible to identify someone by the words they say. This is an abstraction or a limitation of the approach described before. However it's a good idea to test it's performance on our use case.

Chapter 2

State of the Art

In this chapter there will be an extensive explanation to the technologies used throughout the document.

2.1 Biometric factors

Biometrics is the set of characteristics which allows a person to be identified by one or more of its physiological traits. Biometric traits can be used for authentication or identification of a person with a relatively high precision, and usually are more secure than password or token based authentications. Some examples of biometric factors typically used in authentication are: DNA, retina, iris, palm veins, fingerprints and face recognition. There is another subclass of biometric factors also called behavioral patterns, which relates more to the behavior of a person like typing, gait and voice.
Due to its nature biometric factors depends on the technology used to gather them, introducing some noise or errors, for which exists some metrics to evaluate their reliability:

- **False match rate** related to the probability of the system to match incorrectly pattern an input to a non-matching element in the database.
- **False non-match rate** is the probability that the system will reject a correct input.
- **Equal error rate**, the rate at which acceptance and rejection error is equal. The lower the better.
- **Failure to enroll rate**, rate at which the system fails to create a pattern from an input, typical of low quality inputs.
- **False to capture rate**, rate at which the systems fails to detect a biometric input.

2.1.1 Speaker recognition

Identifying a person by his or her voice is an important trait taken for granted in human to human interaction.[17] Everyday we use the voice as an identification pattern to distinguish people we are not seeing, to remember people we've met a long time ago whom face we are not recognizing. Moreover humans ability to identify a person by the voice grows linearly with the familiarity with the person they're talking with, arriving at the point of recognizing someone just from their laugh. On the other hand it is also true that we as humans find sometimes hard to remember a one-time heard voice, or having a tough time recognizing someone from the phone.
Speaker recognition can be accomplished in three ways:

Naive speaker recognition is the process of recognizing familiar voices without any conscious training. This process is used daily by most of the people when interacting to each other.

Forensic speaker recognition is the process used by forensics to identify criminals from their vocal activity. Special listeners are trained to listen phone conversations activity from suspects and compared with the criminal records, trying to find any match between them.

Automatic speaker recognition is the process used by computers to learn about a human and then recognize known voices in audio files.

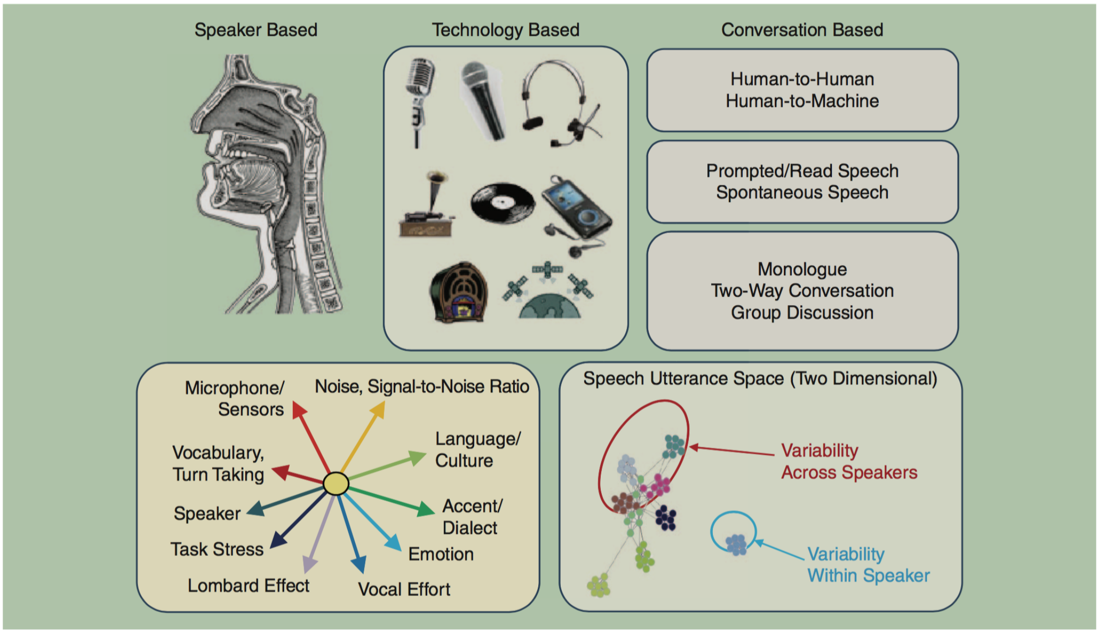
Speaker recognition is a large topic of research which includes two separated taks: speaker identification, speaker verification and speaker diarization. Although we will focus mainly on the identification, speaker verification happens when an unknown claims to be a known person and the task of the system is to verify the correctness of the claim. In speaker identification the task is to identify an unknown speaker from a set of known speakers, using audio samples to find the closest speaker to him or her. Finally, speaker diarization is the process of partitioning an audio stream in clusters relatively to the different speakers in the recording. It is usually combined with speaker recognition and it answers the question: *Who spoke when?*. We define *closed* or *in-set scenario* when all the speakers are known and *out-of-set speaker identification* when the potential input may be out of the predefined speaker group. Moreover speaker recognition can be:

Text dependent when the content of the speech is important to the identification of the person. Different people have different patterns of using words, adding a new trait of distinction between humans. This is usually used to authenticate users in restricted systems or resources.

Text independent when the content doesn't influence the recognition of a known speaker but only the voice traits are used to identify the person.

Human speech is a performance biometric, making it intrinsic different from the majority of the common biometrics (e.g. iris, fingerprints, footprint etc). This is due to the identification made on how something is said rather than what, with a larger degree variability. As can be seen in figure 2.1 The sources of variability in speaker recognition can be partitioned in three major classes: speaker based, conversation based and technology based.

Figure 2.1: Sources of variability in speaker recognition



Source: Speaker Recognition by Machines and Humans

Speaker based variability

These sources of variability depends on the speaker voice traits, for which a range of changes of how the speaker produces a speech and include the following factors:

Situational task stress when the speech is influenced by a task, typically stressful, made by the speaker. This usually happens when someone is speaking meanwhile doing some task such as driving, calling for an emergency. All these tasks includes cognitive or physical stress which introduces variability in how the speaker speaks.

Vocal effort/style happens when the speaker alters his or her voice to react to a noisy situation modulating his or her voice to the situation; furthermore this happens also when singing.

Emotion introduces a bias while speaking due to the human instinct of communicating the speaker emotional state through the tone of his voice.

Physiological source is related to the physiological state of the speaker which can alter the normal voice trait for situations including: illness, cold, aging or medications / surgery.

Disguise is a variability introduced by an intentional alteration of the speaker's voice for exceptional circumstances such as mimicking someone or avoiding detection.

Conversation based variability

These sources of variability are reflected from the conversational situation of the speaker, including the interaction with a human or with a machine.

Human-to-Human speech which may include two or more people interacting and discussing as well as one person addressing an audience. It may also consider different cases involving remote conversation through computer with headphones and a monologue.

Human-to-Machine speech involving an interaction between a human speaker and a machine. This is mostly the case when a human has to adapt his voice to be understood by a computer, for example when using a virtual assistant, namely Siri.

Technology based variability

These sources are introduced by the technology used to acquire the speech, mostly depends on the quality of the input (microphones). Noise is typically introduced from the following channels:

- electromechanical - which includes the transmission channel used to acquire the input.
- environmental - all the noises introduced by the background or from outside all the input components. Some examples may be white noise, background noise from the sun or the environment.
- data quality—duration, sampling rate, recording quality, and audio codec/compression.

Feature Parameters

Each speaker voice has some insight characteristics which identifies uniquely the voice. Individual traits taken from the speaker vocal tract physiology or from the articulation learnt during the time (i.e. dialects). These traits are also called *feature parameters*, defined as a set of measurable and predefined aspects of a speech to be considered for a meaningful comparison between two different voices. Typically multiple features are used to identify a voice, and is more important if we consider the noise introduced by the previous sources, which makes having many features an important aspect to consider. Different voices may have similar or identical values for a specific feature but differ in all the others. As outlined from Nolan[18], typical features has to have the following properties:

- High between-speaker variability and low within-speaker variability
- Resistance to disguise and mimicry
- Robust in transmissions
- High frequency of occurrence in materials
- Easy to extract and measure

2.1.2 Speaker Recognition algorithms

In the following sections we'll provide a brief introduction to the state-of-the-art of speaker recognition methods.

Gaussian Mixture Model Based

A Gaussian Mixture Model (GMM) is a combination of probability density functions (PDFs) used to model multi-variate data. [17] GMM will provide a score evaluating the PDFs at different times, which will be used to calculate the similarity between a speaker's GMM and an unknown speaker. During an identification the system will calculate a GMM for each known speaker, and use it to make comparisons between for which the highest ranking will be selected.
Before GMM there were Vector Quantization(VQ) methods, which used sets of prototypes instead of PDFs to build models for the speakers. However GMM proved to be better because of its probabilistic model being able to relate better to data despite of the more restrictive VQ model. The Model is represented by the sum of the individual PDFs, where a random vector x_n can be modeled by N gaussians with mean vectors μ_g and covariance matrices Σ_g, where g=1,2...N. The PDFs x_n is given by

$$f(x_n|\lambda) = \sum_{g=1}^{N} \pi_g N(x_n|\mu_g, \Sigma_g) \tag{2.1}$$

where π_g is the weight related to the g_{th} component. The model is denoted by $\lambda = \{\pi_g, \mu_g, \Sigma_g | g = 1...N\}$, where the probability of a feature vector given the GMM can be calculated with the above equation. Given a GMM the probability of observing a sequence of feature vectors $X = \{x_n | n \in 1..T\}$ can be computed as

$$p(X|\lambda) = \prod_{n=1}^{T} p(x_n|\lambda) \tag{2.2}$$

GMM-UBM Speaker Verification

Although GMM model based techniques are effective for speaker identification tasks, to achieve speaker recognition a further speaker model is needed. The new approach implies the use of a *Universal Background Model* (UBM) as alternate speaker model. The idea behind UBM is to represent everyone except the speaker itself, essentially a large GMM trained to represent the generic speaker-independent distribution of speech features.

GMM-UBM is an approach combining the previous GMM model adapting it from the UBM using Bayesian adaptation. The model is obtained updating a well trained UBM instead of calculating the maximum likelihood training of the GMM for a speaker. Speaker models obtained from the adaption of a trained UBM are more reliable than GMM trained directly for each speaker.[17]

GMM Supervector SVM

During the years research moved toward having fixed size representations of speech. However this is far from reality where most of utterances used to train models are of different duration lengths. A solution for the problem is the use of *supervectors*, a fixed size vector made concatenating all the variable duration utterances parameters. A GMM supervector is typically obtained by concatenating the GMM mean vectors of a MAP-adapted speaker model.

Supervectors can be used effectively using support vector machines (SVM) for speaker recognition tasks. Basically the supervectors obtained from the training are used as a positive example meanwhile different utterances are used as negative examples for the model. An SVM classifier aim is to separate multidimensional data points obtained from two classes on a hyperplane, and predict the class of an unknown observation from its location on the hyperplane. Given a set of training vectors and labels $(x_n, y_n) forn \in \{1..T\}$ where $x_n \in R^d$ and $y_n \in \{-1, +1\}$. Therefore an SVM goal is to learn a function $f : R^d \rightarrow R$, for which the class label of an unknown vector x can be predicted by:

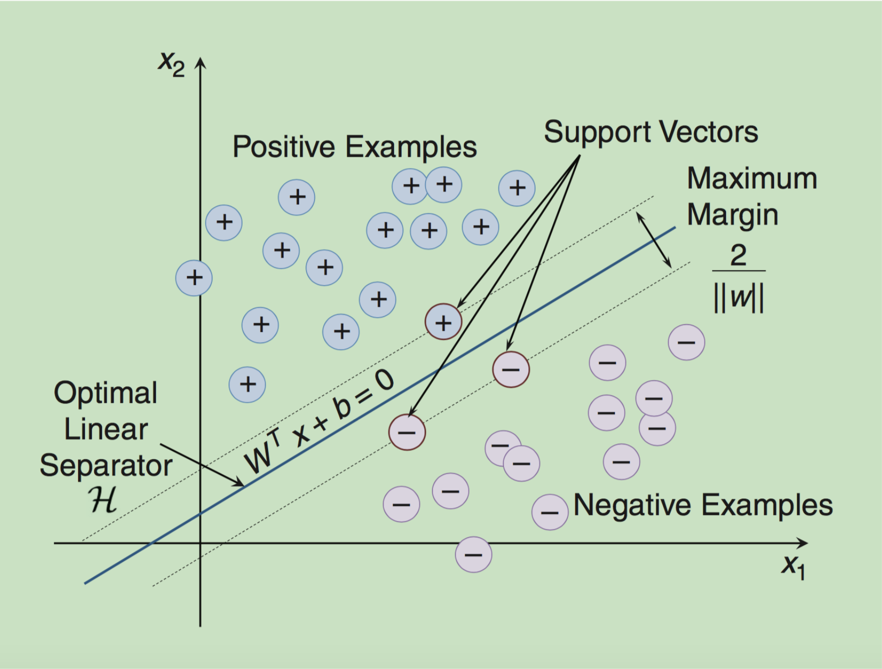
$$I(x) = sign(f(x)) \tag{2.3}$$

A hyperplane H for linearly separable data given by $w^T x + b = 0$ can be a class separator for two classes as:

$$y_n(w^T x_n + b) \geq 1, n = 1...T \tag{2.4}$$

The optimal situation happens when the H provides the maximum margin between classes, the points ling on the margin are also known as support vectors. An example of hyperplane can be seen in figure 2.2. In conclusion this approach is considered very robust due to its ability to combine the effectiveness of adapted GMMs with the discriminating ability of the SVM.

Figure 2.2: Sources of variability in speaker recognition



Source: Speaker Recognition by Machines and Humans

Joint JFA

The Joint Factor Analysis model is built combining both eiganvoice and eiganchannel together, using a MAP adaptation for a single model. The model basic assumption is that both channel and speaker variability lies in the lower dimensional subspaces of GMMs supervector space. These subspaces are modeled by two matrices U and V, where for a random utterance with speaker s and session h a GMM supervector can represented by equation 2.5

$$m_{s,h} = m_0 + Ux_h + Vy_s + Dz_{s,h} \tag{2.5}$$

Joint FA exploits the behavior of speaker features in a variety of conditions learned using FA [17].

i-Vector Approach

An i-Vector uses a set of low-dimensional factors(w) to represent the conversations side. These factors controls an eigan-dimension of the total variability matrix (T) and are also called *i-vectors.*

$$s = m + Tw \tag{2.6}$$

Where

- s is the conversation side supervector
- T the total variability matrix
- w is the i-vector

Unlike JFA or other factor analysis models, i-vector does not consider any distinction between speaker and channel but rather consider them as a dimensional reduction of a GMM supervector. I-Vectors effectively summarize utterances allowing compensating methods that were unpractical before with large supervectors.

2.2 Internet of Things

Internet of Things, better known as IoT, is the world of interconnected devices connected to the real world and to the Internet. These devices has a very broad definition, from the smart sensors in a power plant to the smart fridge in a house. What connects these devices is their ability to be connected with the world, sharing, with the needed restrictions, their work. This opened the door to a new revolution in the IT sector, prompting new opportunities for developers, entrepreneurs and end users. Some of the most relevant endings of these technologies are *Smart Houses*, *Smart Cities*, Cars etc.

2.2.1 Ecosystems

An ecosystems is by definition *a system, or a group of interconnected elements, formed by the interaction of a community of organisms with their environment.*, in our case a set of smart devices connected between them. The IoT world is evolving from distinct single entities to more evolved clusters of devices which can communicate more easily between each other.These are usually are made by companies using specific protocols or standards to simplify the connection between their products, but at the same time closing it to the others. The list of products related to the *Internet of Things* is too broad for being even listed, due to the highly expanding sector related to smart devices. However we will consider the most common ecosystems which can be easily found in a common home, meaning services for: Smart Lights, Cameras, Thermostats etc. Luckily most of them are being bought and integrated in bigger environments run by famous companies such as *Google* or *Apple*. The reason for which we try to restrict the integration with these ecosystems is mainly practical, because they're the most common and they do offer simulators for testing, and because they suits perfectly our scenario.

Google - Nest

Nest[19] is a home automation producer of smart devices which ranges from their famous Smart Thermostat to the locking system, from the washing machine to the light system, from the *Dropcamera* to the Hi-Fi Sound system,and these are just some examples. The number of components supported by this company is wide, which makes it a good choice to support in our scenario, also because of the high cost of these devices. The main added value given by these accessories is related to two main points: saving money with a smart consumption of electricity and the ability to remotely control the house with an easy to use application. Besides the commercial value of *Nest*, the platform offers many tools for developers to interact with their platform using their Cloud service. Third party developers are encouraged, with some limitations, to use their Cloud service which offers some *RESTful* APIs to gather access to the remote devices. The remote access through APIs unlinks developers from platform dependent libraries (see later HomeKit) that restricts the use or the integration with a specific technology. Furthermore it is possible to access directly *Nest* devices with their *Nest Wave* which allows direct communication with non-branded devices using two different communication protocols, mainly *802.11* standards.

Apple - HomeKit

HomeKit[20] is a very similar ecosystem to the one described before, producing or supporting home automation devices. *HomeKit* relies on the large network of Apple devices, making it an environment to be considered even if it is relatively new on the market. The key point of *HomeKit* is the easy integration with Apple devices, fitting almost perfectly wherever there was an already existing Apple ecosystem. Developers are allowed to take control of the devices only through iOS applications, restricting the possibilities of integration with different ecosystems or technologies. Moreover it ties the developers to their technology making it really hard to be adopted in a different system. However as we'll see later it is possible to bypass the problem using the microservice architecture to make the system independent from specific technologies.

Samsung - SmartThings

SmartThings[21] is another ecosystem from *Samsungs* with many similarities with the former producers.*SmartThings* focuses on four main Smart devices categories: *Security*,*Monitoring*, *Lighting and Energy* and *Convenience and Entertainment. SmartThings* as the previous ecosystem does not allow a direct interaction with the smart devices, but offers a developer-friendly interface to their Cloud Service. However, as *HomeKit* it is tied to a technology, making it harder for different technology stacks to access.

2.3 The Microservice Architectural Style

SOA vs Microservices

In short, the *microservice architectural style* is an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API. These services are built around business capabilities and independently deployable by fully automated deployment machinery. There is a bare minimum of centralized management of these services, which may be written in different programming languages and use different data storage technologies.[7]

There is a close link between the *microservice architecture* and the *service oriented architecture*, thus due to their nature the community classified microservices as a subclass of the service oriented architecture. *Don Box* of Microsoft described the Service-Oriented paradigm with the following four principles [8]

1. Boundaries are explicit

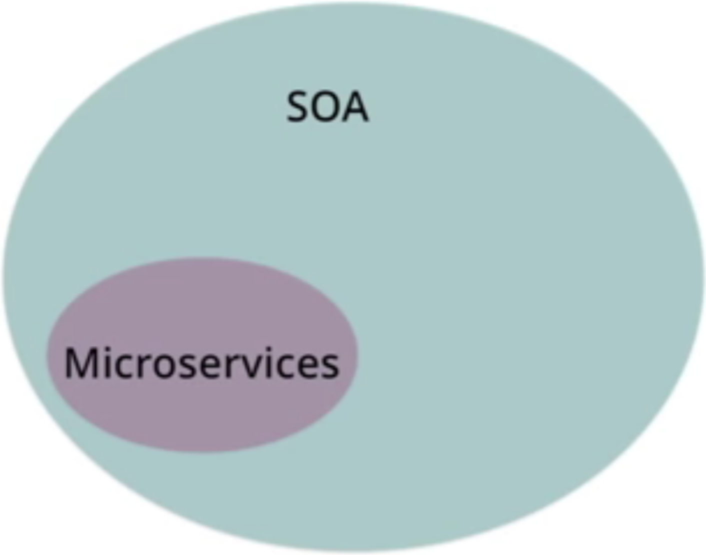
2. Services are autonomous

3. Services share schema and contracts, not class

4. Service compatibility is based on policy

Microservices fullfills the first two requirements, with a very strong focus on the second principle. However the functionalities are very frequently exposed using a *RESTful* interface, which doesn't expose any contract nor schema. Furthermore microservices holds another subtle difference related to their scope, where a microservice serves as a service inside its application meanwhile typical SOA services serves a broader scope and *can be* part of the same application. The difference between the two concepts is very subtle, and it wouldn't be impossible for them to be the same in certain situations. *Bob Ruhbart* from Oracle described shortly the difference: *Microservices are the kind of SOA we have been talking about for the last decade. Microservices must be independently deployable, whereas SOA services are often implemented in deployment monoliths. Classic SOA is more platform driven, so microservices offer more choices in all dimensions.*[10]

Figure 2.3: Graphic illustration of microservices and SOA



2.3.1 Internal integration

Typically when a new component has to be added to an existing project the approach consists in the development of a library which has the logic to deal with the new module to be supported. Subsequently the component will become part of the project itself, with its needs to be updated throughout time. However when the number of components to integrate increases it will affect the size of the project and very likely its performances. In our case the component to be integrated will be the ecosystem driver, having a set of libraries which are capable of interacting with the remote APIs or with the direct wired connections. As we'll see in the next paragraph this is a typical monolith approach, meanwhile for this situation would be better to use a microservice approach.

2.3.2 Integration as a Service

Considering the *Microservice* architectural pattern we can decompose the above situation creating dedicated services capable of handling the required business logic to interact with an external system. This approach is also called *Componentization via Services*[7], where a component is defined as a *unit of software that is independently replaceable and upgradeable.* It is important to distinguish between *libraries* and *services*: the latter uses out-of-service components to communicate, mainly HTTP requests or remote procedure calls when libraries uses instead mechanisms like in-memory calls. The main advantage to build components as services instead of libraries is the possibility to deploy them individually without the need to redeploy the whole system. If a library is modified or removed the whole system will need to be redeployed, which in most of the cases it is converted in a loss of money and time. That's not the case if the system is composed by many independent systems, where only the changed service will need a new deployment. However this is not always true, there will be some circumstances where it will be necessary anyway to deploy again the whole system, but the aim of this approach is to reduce the number of these necessities.

2.3.3 Other benefits

Apart from the benefits already listed formerly, there are other benefits introduced by adopting the microservice architecture:

Heterogeneity between technologies Structuring the system as a set of services frees us from the limitations of a singular technology allowing us to adopt different frameworks for different tasks. This benefits on the possible optimization that can be achieved using the right technology for the right task. Furthermore this removes completely the problem of creating adapters for different technologies to integrate in the system if they do not exist. This is also called *Decentralized Governance.*

Evolutionary Design is a concept popularized by the *Extreme Programming*, where the system is continuously evolving during the phases of it's development. This key feature makes possible to evolve adopting a microservice architecture while keeping the old legacy monolith system, well tested and functional, without rewriting the whole system.

Designed for failure Building a system made of services instead of components leads the developers to take more effort in considering failures. Developers have to take in consideration the possible failure while reaching the service, and prevent the system to crash and handle the situation in the most gentle way. On the other side, this approach introduces more overhead to handle the possible situations.

2.3.4 Microservice Oriented Internet of Things

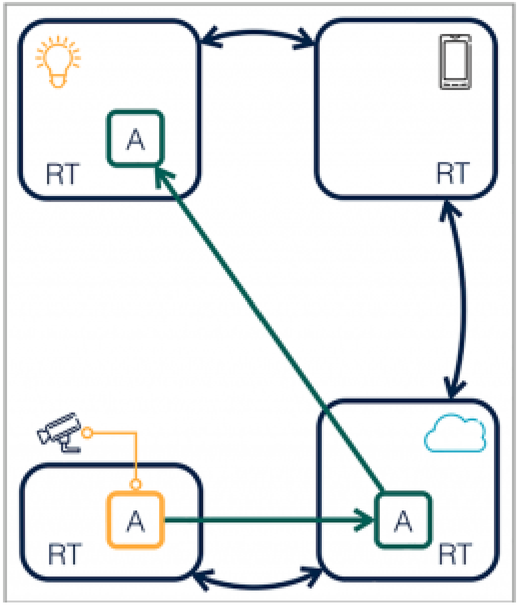
IoT with their advent brought up some, not yet resolved, challenges which fits very well with the Microservice architecture. *Interoperability* is one of these challenges aimed to be solved. Microservices *decentralized governance* feature, which allows the use of different technologies inside the same systems holds the key for solving the interoperability problem. Key IoT devices vendors realizes that if their products supports multiple interoperability standard they're likely to be adopted also by their competitors (e.g. Nest products are available on HomeKit[3]). Microservices here can be used as isolated adapters for the many different technologies, facing the spreading fragmentation of communication protocols between machines.
Microservices are by nature dynamic and reactive, which makes them a good choice for an environment where new technologies are frequently added or modified.

2.4 Calvin

Calvin is an Open-Source distributed framework designed mainly, but not only, for IoT systems. The key point of *Calvin* is the "Distributed cloud for IoT", meaning running the code where it best suits the performance needs, a crucial aspect since we are dealing with low-resource components. We will use *Calvin* for the whole course of the document, referring to it as the main "system", also used to integrate with the others.
Calvin has the ability to integrate new components without replacing them by writing proxies for the specific hardware to support. Proxies are actors written to handle communication with the legacy system. Such a proxy actor handles the task of converting data from the application into messages or requests the old system can handle, and converting the response back into tokens the Calvin application understands.[14]

Figure 2.4 shows the interaction of native Actors and Actors which acts as a proxy for

Figure 2.4: Proxies interaction



Source: Ericsson Blog

legacy components, in this example the camera.

Calvin is built upon the well-established actor model, using a methodology often referred to as dataflow programming[11], and its life cycle can be summarized in four, well-distinct, phases: *Describe*, *Connect*, *Manage* and *Deploy*.

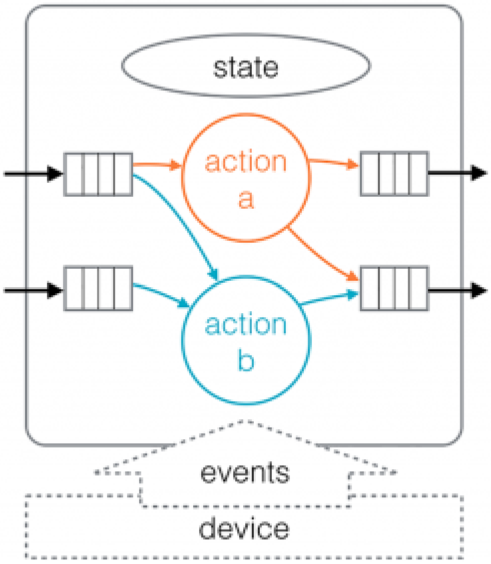
2.4.1 Describe

The smallest functional units in *Calvin* is an Actor. Actors do not share nor state or behavior, encapsulating all the logic. The key point of Actors in Calvin is their reactivity: they react to external events or when receiving inputs. Actors communicates only using data through ports, has to be defined before deploying the system. This way is possible to describe the possible interactions that the actor may have when connecting it to others. Having a non-shared internal state allow the Actor to be serialized and moved to another running machine or to be backed up if the machine crashes. However this is partly true because it may be tricky to serialize Actors with a very complex internal state.

2.4.2 Connect

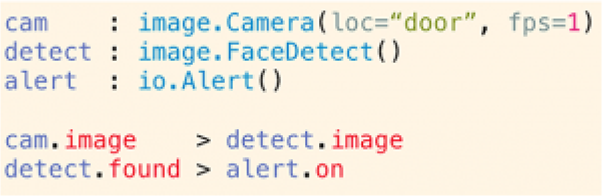
After describing the many functionalities provided by the various Actors in the system we need to tie them to build applications. *Calvin* offer a scripting language, named *Calvin-Script*, to describe the various connections between actors and their input and output

Figure 2.5: Describe Phase



Source: Ericsson Blog

Figure 2.6: CalvinScript Example



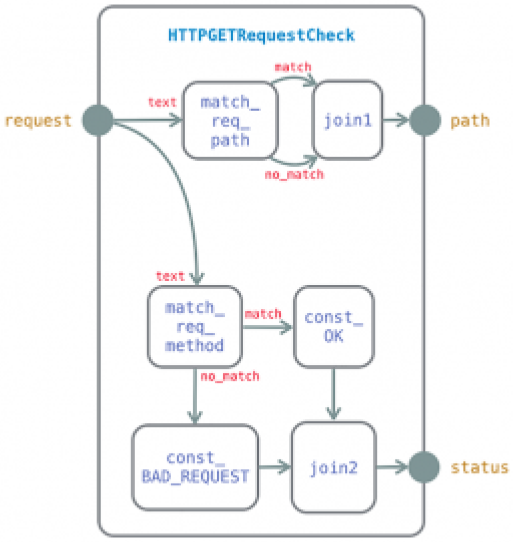
```
cam    : image.Camera(loc="door", fps=1)
detect : image.FaceDetect()
alert  : io.Alert()

cam.image    > detect.image
detect.found > alert.on
```

Source: Ericsson Blog

ports. In figure 2.6 there is an example of a script for detecting faces in an image. The first part is relative to the actors declaration and initialization, giving a clear description of which actors will be playing in the current environment. The second part describes the links between each actors, structuring the flow of the process. In this case the actors in the system are 3: the *Camera*, the *Detector* and the *Alert*. The flow of the process is relatively simple in this case, and it is structured as follows: the *Camera* takes a picture, which will be passed to the *Detector*. Subsequently the *Detector* will send its result to the *Alert* actor, which will possibly fire an "alarm" if he detects any human face inside the picture.
Figure 2.7 shows a more complex interaction between actors.

Figure 2.7: Actors interaction



Source: Ericsson Blog

2.4.3 Deploy

Each Actor has different resource requirements needs to be satisfied in order to be deployed successfully. For example, referring to the previous scenario, an actor using a camera needs to be deployed on a system where there actually is a camera. These are also called *hard* or *unconditional requirements*, which determines the possibility to instantiate or migrate the actor on a machine.

On the other side there are also *non-functional requirements*, describing where an actor would suit best for its task. Always referring to the former example, when applying face detection it would be better to perform this task on a more performing machine compared to a low-resource machine, like a *Raspberry Pi*.

However at the moment *Calvin* supports only *static deployment*, needing the user to define where to deploy and execute the actors.

2.4.4 Manage

When the whole runtime is running it is more than needed to have a tool to keep track and monitor the activities in the system. The runtime can be queried to retrieve informations about the actors and locate them. It is also possible to track the actors firings, on which ports and step by step execution. These are mainly debugging tools, though it is also true that this is a recent framework and many functionalities are already in development.

2.4.5 The relationship between the Actor Model and Microservices

Actors are isolated, single-threaded components that encapsulates both state and behavior. Typically actors communicates using lightweight direct messaging systems, for example to receive or return inputs/parameters. Microservices are very similar to Actors in many of their key aspects, such as isolation, encapsulation and lightweight messaging, though usually microservices uses RESTful interfaces. There is some discussion in the community whether some argues that actors are actually microservices themselves [12] meanwhile others defines actors as a subset of microservices[13]. This however heavily depends on the actors framework adopted for the situation, which in our case an Actor can not be compared to a microservice due to an insight limitation: *Calvin Actors can not use different technologies*, at the moment.

2.4.6 Calvin's Three Tier architecture

Calvin is structured with a three tier architecture: the actor layer, the system layer and the runtime layer. Each of them is isolated and can communicate with each other only passing values through the various layers. The architecture is summarized in picture 2.8

Actor Layer The actor layer is the upper layer which exposes all the actors and their functionalities. Actors can used only the functionalities offered by the lower level calvinsys with no direct access to the runtime system.

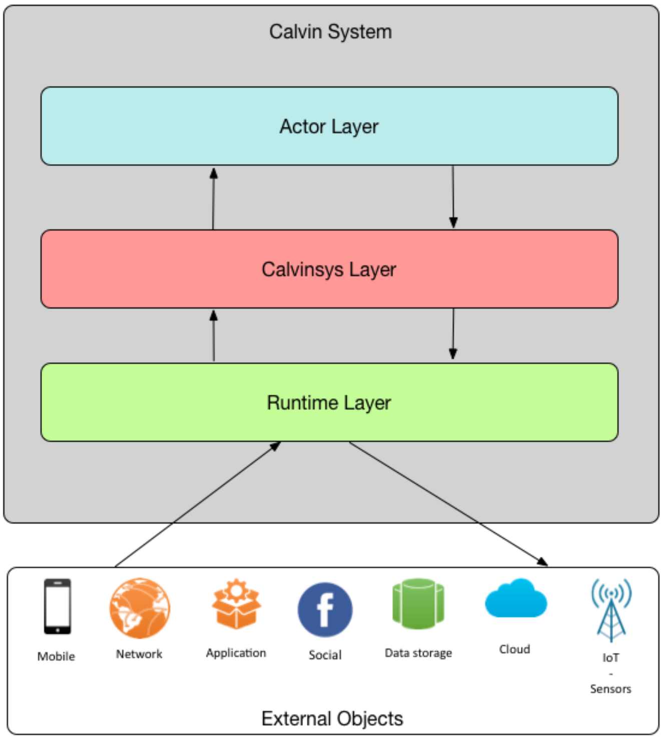
Calvinsys Layer Calvinsys is considered the middleware, where all the runtime logic is exposed through high level functions and objects. This layers wraps and uses the functionalities offered by the runtime layer, offering tools for the development of advanced actors.

Runtime Layer the runtime layer is considered the core, in this layer are stored all the low level implementation.In this layer can be found detailed implementations of how to communicate with different kinds of sensors, an asynchronous implementation of HTTP calls for Calvin and etc. Here's the core of all the functionalities adopted by the higher level objects such Actors or middleware libraries.

2.4.7 Anatomy of an Actor

As previously introduced Actors are abstract reactive entities which exposes inbound and outbound connections. In contemplation of *reactivity* all the actions taken from an actor must be simple, fast and short. Nonetheless this optimal situation is hardly achievable in

Figure 2.8: Calvin Architecture



real life, it's usually more common to have blocking calls to services in the runtime which require on average more then 1-2 seconds. This limitation can be solved simply moving the logic in the runtime towards an asynchronous approach for long time tasks. Typical tasks requiring asynchronous interaction can be reading data from a sensor, sending an HTTP request or authenticating into any service, the list is longer but can summarize in most of interaction with out of the system objects.
An actor anatomy can summarized in the following parts:

- Declaration of the connections
- Initialization and setup
- Migration
- Actions, Guards and Conditions

Declaring the connections

All the connections of an actor has to be declared a priori, and hence be all used (or allocated) when the actor is used. The declaration is done simply adding under a comment the names of the many connections and their type: either input/output or both. Here is a simple example of an actor for HTTP GET requests:

```
Input:
  URL : URL to get
  params : Optional parameters to request as a JSON dictionary
  header: JSON dictionary with headers to include in request
Output:
  status: 200/404/...
  header: JSON dictionary of incoming headers
  data : body of request
```

In this example we can see the 3 input ports: URL, params and header, which will all be needed to be set in order to fire one of the actions, depending on the condition. These inputs can be used as a condition parameter using `action_input['var_name']` and `action_output['var_name']` for outputs.
However it is very likely that we won't need all these inputs or outputs, specially if they are optional like `params`. For this reason there exists two special actors which will connect to these ports without generating tokens or consuming them without doing anything (namely Void and Terminator).

Initial phase

The initial phase is denoted by the decorator `@manage['inputs']`, where all the variables are set and the required libraries are loaded. It is possible to pass inputs at instantiation time using manage and bypassing the input ports, though this approach is discouraged and advised only for particular situations. Furthermore in this phase all the components from the *calvinsys* will be loaded for a consequent use during the actor actions.

Migration

One of *Calvin's* key feature is the simplicity of migrating actors through the different nodes in a cloud system. Migration of the object is achieved using a special function `migrate(self):` to which can be passed the current actor object, and it will be serialized. Migrating implies serializing all the current variables which composes the actor's state to be sent, though problems may arise if these variable contains complex objects such as connections (e.g. DataBase connection, TCP connection etc). When the state will be sent to another machine, this will setup the actor passing from the former phase, hence creating a new (almost) identical actor.

Actions, Guards and Conditions

The core of an actor are the actions it can perform, but these needs some conditions to bet met before being fired (e.g. having a variable set to `true`). Most of actions are composed of three key elements:

Condition is a mandatory decorator which describes the inbound and outbound connections for the action. Although the name suggests it to be the condition statement to for firing an action, condition is related to the input and output condition of the action. An example of condition for a camera actor can be `@condition(action_input=['trigger'], action_output=['image'])`. In this case our actor when will receive a token on the `trigger` port will enable the action, but the next decorator will be the definitive check to pass to the action body.

Guard is an optional decorator, for checking some conditions before passing to the body of the action. Guard is more related to the internal state of the actor instead of the connected ports, thus it's the real condition check to fire the action. Guard can be used to see if a variable has been set or if its value is correct. For example if we want to check that an actor has completed an asynchronous task we can use a guard similar to this one: `@guard(lambda self: self.camera.picture_taken)`. In this case *Calvin* will perform an evaluation of one the actor internal objects: it the value is set to `true`, it means that the picture has been taken and its stream has been saved on the disk (i.e. the file exists). This allow the system to prevent firing actions when the actor may not be ready for it even though the firing connections are set.

Action The action is a simple function which has to take as parameter the inbound values in the condition statement, and return an `ActionResult` object with the expected output bounded to the outbound values.

Moreover inside the actor the actions has to be listed by priority, to avoid ambiguity when choosing with the correct action to fire.

Chapter 3

Detection and Integration

3.1 Preface

The major problem faced during this work is related to the detection of intruders while surveilling the house. Modern security systems applies different techniques to detect intrusions using many different sensors such as leap motion detection, face detection or sound analysis. They have been working fine during the last years, however thieves improves bypassing or partially avoiding these technique making the latter less effective.
Therefore most of the work will be done investigating and testing innovative and advanced techniques to deal with intrusions detection.

3.2 Fingerprinting a Human

During the time there has been many approaches to recognize a "Person" using some of its peculiarities: software side with passwords or access phrases, which should identify a single person; biometrics with finger prints, retina scans and voice recognition.
Usually such controls where available only with expensive solutions (and some still are, e.g. retina scanners), however with the introduction of IoT their price dropped allowing them to be sold for a reasonable price. These controls are also called **behavioral biometrics**, the set of behavioral traits which identifies an individual person. Usually these traits are used for authentication when accessing a restricted system, however we'll consider these to recognize unknown entities accessing our house.
These systems are mainly needed to prevent accidental triggering of the alarm, to be more precise when listening or watching, not detecting just something but a real intrusion.

3.2.1 Behavioral biometrics

Detecting if someone in the house is someone familiar is not an easy task, most of the approaches are precise in optimal conditions, much less when the situation is non optimal (low light, low resolution etc). This impacts negatively their efficiency when applied in a real life scenario, where usually the conditions are far from the optimal situation. For example facial recognition works in very limited cases, like the systems used on smartphones to unlock the device, has been reported to not be reliable enough to be the main unlocking system. The situation is also similar for voice recognition or speech to text systems, which are improving during the time and mainly **during the usage**. *Siri*, one of the most famous personal assistant had some minor troubles at the beginning being used to the different users voices, accents and languages. *Siri* uses different voice recognition algorithms to categorize your voice by learning your dictation and your accent, with good results over time.

3.3 In House Speaker Recognition

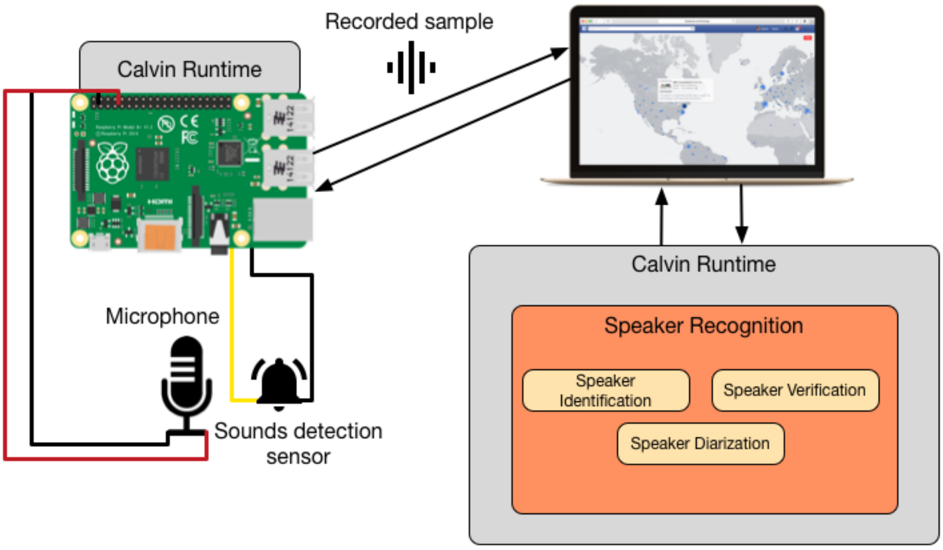
A house is an ideal place where to place a speaker recognition system, since there are not many noise sources, ensuring a good quality input for microphones. Hence a good surveillance system with many microphones around the house should be able to detect any external voice during an intrusion. In the following paragraphs we'll show how we implemented such a system using some of the state-of-the art algorithm with *Calvin* to improve the security in the house.

3.3.1 Speaker recognition and Calvin

Machine learning tasks are typically cpu and time consuming tasks, which are not supposed to run on low-resource machines like a Raspberry Pi. However we need to access physically some sensors to gather the input and not many computers offer direct GPIO access, making devices like a Raspberry Pi a must when interfacing with hardware. Despite of this limitation *Calvin* offers the solution to this problem when adopting the philosophy of *"run it where you need"*. As said in the previous chapter Calvin actors can run on different machines, and communicate through the network. As can be seen in figure 3.1, in the basic configuration we need two machines: one interfacing with the hardware, mainly a microphone and a sound detector, and a more powerful machine.

In more detail, in our architecture there will be 2 actors controlling the input from the two sensors, and at least 3 in the other machine, each of them taking care of a different

Figure 3.1: Spaker Recognition Architecture with Calvin



task, namely identification, diarization and verification.

3.3.2 Actor Structure and Interaction

The interaction between actors in this situation can be summarized as a chain of events triggered by the detection of a sound to the detection of a person, either familiar or unknown. In our proposed architecture an Actor linked to the sound detector sensor will continue listening, and when something is detected it will generate a token for the microphone Actor to switch to the recording state for a finite period of time. However it is not possible to stream in real time the recording to another machine without overcomplicating the system, for which we proposed to split the original audio recording in small pieces to be sent through the network. This approach has some theoretical benefits:

- Dividing the record in small units make it faster to detect someone, since the verification phase will require less time.
- Small audio recorded files are easier to send instead of a single big one.

Despite of the benefits, in the following chapter we will try to prove the effectiveness of this approach, and the best duration of time for each time slot.
After recording the audio slice the Actor will have to send it to the identification /diarization/ verification actor residing on a different machine. However sending binary objects through the network may be tricky without working directly with sockets, in our solution

we decided to use the *base64* encoding and envelope it into a JSON object. The object will be decoded and stored as a binary object and subsequently the machine learning actor will execute its algorithm on it.

3.3.3 Hardware Interacting Actors

In this section we'll go in the details of the low level actors handling the resources such as sound detection and microphone, describing their internal logic and actions.

Sound Detection Actor

The simplest actor is the one dealing with the sound detection. It's state works as a simple switch turned on by the change of state of the sensors. This actor does accept only an input, the trigger for reading the sensor value. This allows us to decide at which rate reading the sensor without leaving the read rate to the internal scheduler. The actor generates only a token each time the value read by the sensor is equal to 1 (i.e. a sound is detected). Therefore the conditions of the Actor will be like:

```
Condition for receiving the input
@condition(action_input=['trigger'])

Condition for returning the output
@condition(action_output=['token'])
```

These conditions will bound the input and outputs to the correct functions. Furthermore, to clarify the internal state change of the actor here is shown how the control on the sensor reading using the *Calvin* `@guard` statement:

```
@guard(lambda self: self['sensor'].read_state() == 1 and self.trigger)
```

The sensor can be accessed through a managed object handled by the *calvinsys* layer, hiding the implementation details. Moreover the `self.trigger` secondary condition requires the Actor to be in possession of a valid token, which will be consumed after each reading.

Recording Actor

The recording actor will have some more complicated logic having to deal with the audio sampling. The Actor will be bounded to the former actor to receive the token and start

the recording process. As said previously we decided to take a different approach, instead of a single recording we will sample it in n segments of the same duration. The Actor will therefore change its state in order to generate the needed amount of audio recordings to be sent to the speaker recognition Actor(s).
Here are the conditions applied to the actor in order to fire the many actions:

```
Initialize the actor with a number of samples and their duration
@manage(['n_samples','duration'])
This is a one time initialization, without bounding these parameters
to any inbound port.

The condition for setting the actor as active, coming from the
sound detector
@condition(action_input['token'])

The outbound connection is a bas64 encoded audio recording loaded
into a json object
@condition(action_output['wav'])
```

However the former skeleton doesn't explain enough about the internal structure of the Actor, for which the more specific guard conditions explains it in more detail.

```
This guard condition starts the recording process of an audio sample,
with three conditions to match:
token available, the number of recordings done and the microphone state

@guard(lambda self: self.token and self.token_used > 0 and
                                   not self['microphone'].recording())
Here the actor checks if the there is any wav file to dispatch
@guard(lambda self: self.wav)
```

It is necessary to clarify the behavior of some internal variables:

- token - is a variable set to True each time a token is received from the previous actor. It gets voided when the actor dispatches all the n audio recordings.

- token_used - an internal variable initialized to n each time the actor receives a token and decremented each time the actor sends an audio file to the speaker recognition actor.
- wav - is a simple flag representing the availability of a wav file to dispatch
- microphone - is the physical sensor managed from the *calvinsys* layer.

3.3.4 Speaker Recognition Actors

In this section we will go further with the Actors specialized in the *Speaker Recognition* task, namely: Speaker identification and speaker diarization.

Speaker Diarization

As introduced in the previous chapter, speaker diarization is the process of finding the various speakers in an audio recording. The Actor supports only two functions: extracting the speakers and adding models to use in the identification phase. The diarization works in two phases: first the system splits the recording in many clusters based on who is speaking, and subsequently it tries to identify the voices from the models in the database.
The Actor accepts the following inputs:

- wav - The base64 encoded audio recording from an audio source (i.e. the former microphone actor)
- operation - The operation to perform on the file, either extracting speakers or adding a model to the database.
- model name - An unique identifier for the new model to be added in the database

Therefore the output connections will be:

- number of matches - the number of different speakers in the audio recording
- matches - the models for which there was a positive match in the database, unknown otherwise.

The following conditions apply in order to maintain a proper flow in the execution:

```
These conditions only apply when receiving and returning
values:
@condition(action_input=['operation'])
```

```
@condition(action_input=['wav'])
@condition(action_input=['model_name'])
@condition(action_output=['n_people','names'])
```

However these conditions doesn't express the real logic beneath the Actor, which is handled by the `@guard` conditions. Here follows the most important guard conditions.

```
This guard allow us to set the operation only when it's unset,
restricting the set of possible choices for the operation.
@guard(lambda self, operation: self.operation ==
    None and (operation == 'extract' or operation == 'add_model')

The following guards conditions are used for the two steps
asynchronous recognition. The first phase start the process
and needs some parameters to be set properly:
the operation must be extract, the wav file has to be set
and extracted means we did not pass already this phase.
@guard(lambda self: self.operation == 'extract' and self.wav
    is not None and not self.extracted)

Here the same applies for the wav file, meanwhile
one of the conditions to enter the function is the end of
the previous process, indicated by the internal state
._in_progress.
@guard(lambda self: self.wav is not None and
    self.recognizer._in_progress == None and self.extracted)
```

In conclusion, the Actor's purpose is to identify the different speakers and if they match with someone in the database return their names.

Speaker Verification and Identification

Due to its very subtle difference, we have decided to encapsulate verification and identification in a single actor. Speaker Identification is applied when no model to identify is given, meanwhile verification is executed when a model to match is available. The structure of the Actor is very similar to the one proposed in the previous: the actor accepts as input a wav file (base64 encoded) and a model name (in this case is optional), and an operation parameter for adding models to the database. Therefore the input are equal to the previous actor, though this time operation may be only restricted to *recognize* and *add model*. However the outbound connection of the Actor will be different, this time the actor will return the likelihood of the match, and the model for which there was a match (if it is identification).

3.4 Expanding the system

One of the biggest limitations of nowadays security system is their isolation when installed. Most of devices are stand alone solutions incapable of extending their functionalities to pre existing devices. If we want to expand *Calvin* to be adopted by others, including users and developers, we need to be able to extend its functionalities with the ability to communicate with different other systems.

3.5 The Architecture

In this section we will go deep in the details of the architecture of the solutions we propose.

3.5.1 Accessing different systems

Accessing different systems is a problem of technology interoperability between different technologies running on architecturally different hardwares. Most of the times *IoT* devices runs on low limited power devices such as an *Arduino* or a *Raspberry Pi*, both supporting different system languages (the first is limited to a subset of *C++* meanwhile the second one supports everything that can run under a normal linux distribution, but usually it's *Python*). However in our scenario the devices we consider are on a different layer of abstraction, we do target ecosystems and not single independent devices. **Ecosystems** has to provide an access to developers through a point of access, **REST api** for *Nest* and the **HomeKit.framework** for accessing *HomeKit* devices. The limitations are not always the same, when using *Nest* it's easier to communicate with it using a wrapper for their own api, which exists in many libraries, including *Python*. Nonetheless the situation with *HomeKit* is much more strict, having their ecosystem restricted only to their technology, *Objective-C* or *Swift*, which runs only on *Apple* branded device. Furthermore, the limitations are even more strict due to the limitation of the framework only for *iOS*, at the moment at least. Usually the latter is the common case, which implies the need of a different solution then "just wrapping" the apis in the code and plug them into the project. In our case *Calvin's* system is fully written in *Python*, though it have many extension for a wide range of technologies, communicating with *HomeKit* is still hard to achieve. In the following sections we'll describe the approaches taken to solve the problem.

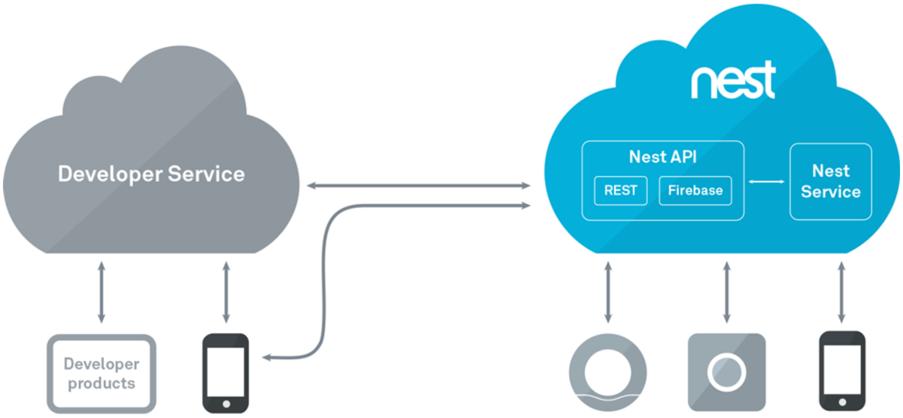
3.6 Accessing Nest

As said in the previous section *Nest* has some advantages due to their cloud apis, which makes it easier to reach the connected devices.

3.6.1 Nest API Model

Nest offers a cloud API near real time, based on a subscription, that allows users to build products accessing data on their devices, with the relative ability to read and write data. With the *Nest* cloud each element is identified as a resource and accessible through a unique address, called "data location". Furthermore, the cloud system also offers a *RESTful* service to access these resources, but only allowing GET and PUT operations with a call limit to prevent overuse. There is also a **Streaming** feature for REST, which allows our application to receive updates in real-time or to stream from a device, namely camera, since web-sockets are not supported. However it applies the same rule as before, with some limits in the data usage to prevent abuse.

Figure 3.2: Nest Cloud architecture



Source: Nest Inc

As can be seen in figure 3.2, the only way to access *Nest* devices at the moment is only through their cloud system. However this will add many layers of overhead and delay due to the various layers, including calling services external to our network. However, in the upcoming future it will be also possible to use *Nest Wave*, a direct device to device communication through protocols such as BLE or Wi-Fi, moving the integration to a higher lever of quality.

The cloud functions of *Nest* are limited to a restricted set of devices: Thermostat, Protect (alarm), Camera and the Home. Here are the functionalities offered by each of the products, with a brief description:

Nest Thermostat

Thermostat, as the name suggests, is a smart thermostat ,which can also be remotely controlled, with some energy saving settings to help the user avoiding unnecessary heating expenses.
Its remote functionalities are:

- Read the current temperature
- Read or set a target temperature
- Set the fan timer
- Read or set the temperature mode
- See humidity values
- View online status and last connection information
- Read structure name and device location in the home

Nest Protect

Protect is a location based alarm for dangerous gas leaks, such as Carbon monoxide *CO*, or smoke in case of fire.

- View CO or smoke status
- View battery health state
- View last manual test status and timestamp for last manual test
- View online status and last connection information
- Structure name and device location in the home

Nest Camera

Camera, previously known as Dropcamera, is one of they key components of the *Nest* suite. It provides screen capture, audio and video streaming and the other typical features offered by smart cameras.

- View camera online status or mic status
- View or change streaming status (turn video streaming on/off)
- View device name and where identifier
- View last online status change (last online/offline change)
- View subscription status (enrolled/not enrolled)
- Learn more about Nest Aware with Video History subscriptions >
- View deep links to the live camera feed in the Nest app (iOS, Android) or on the web at home.nest.com
- View content related to the last event that triggered a notification, including:
  Sound or motion event detected
  Event start/stop times
- Deep links to image and gif files
- Structure name and device location in the home

3.6.2 Nest with Calvin

One of the key aspects of *Calvin's* Actors is their abstraction to allow their reuse in different situations. However, due to the different nature of the ecosystem, specially *Nest* and *HomeKit* the actors can't be used for the same purpose. *Nest* structure can be decomposed in two big sections: Structures and Devices. Structure are objects to keep grouped together many devices, placing them in different part of the house, but still not the most important part in our scenario.

A device is an abstraction of all the smart products offered by Nest, but mainly the one we have listed in the latter paragraph.

Device Abstraction

We have abstracted a device in its most basic functions: setting and getting properties. This simplification allows us to follow the best practices for the development of *Calvin*, allowing a possible reuse of the actor. Here is the pseudocode for the high-level functions definition.

```
getproperty: in device_id, property_name; out property_value
setproperty: in device_id, property_name, value ; out True or False
```

However going deeper into the actor architecture we have to define some more constraints, like the events which trigger the execution and the tokens for which returning the results. All incoming port of an actor has to be linked, we can not have an unbounded port, both input or output. Here's a list of our actor ports:

```
Inputs:
    device: the device identifier
    operation: string representing the operation to do (get or set)
    property_name: the property to be get/set
    value: the value to be used to set the property
Outputs:
    result: the property value
```

To simplify the work, we reduced the inbound port for the operation to only one, instead of having a token for a *get* or *set* operation. This task is completed using the `@condition` to declare all the elements which are needed to fire and will be used in a particular action. A simple condition for an action is the following example:

```
@condition(action_input=['operation'], action_output=[])
```

This condition requires the token operation to be bounded, and it will use it in its body function. Since no `action_output` is defined, the operation does not return any value.

Once all the inbound ports are connected, i.e. they have a token, the system will check the `@guard` condition, which acts as a conditional control: it will continue only if the condition is satisfied. An example guard condition can be written as:

```
@guard(lambda self: self.nest._in_progress is None and self.operation == "set")
```

This guard condition is used to enter the function only when the *nest* object is not working on an asynchronous task and the operation has to be the *set* operation.
One of the constraints that an actor has to follow is the **reactivity**, which implies non-blocking calls in the body of its functions. However, as in our case, the actor will have to make some calls to the *Nest* cloud, which is a blocking call. This is reached using a two step asynchronous system: first the value is acquired and the actor will delegate the request to a different thread; second when the delegate terminates it will modify a flag to mark its completion, at that point the actor will return the value obtained from the nest cloud. Asynchronous calls are needed since the scheduler loop can't be blocked during the execution otherwise it would break the reactivity of the other actors.

3.6.3 A Microservice Oriented approach

Despite of the correctness of the former solution, it does however contains some flaws in the architecture. The solution works only under the following circumstances:

1. Libraries for connecting to a different ecosystem are provided in the same language (e.g. *Python*).

2. We assume the actor does never crash, and if it does the system may crash subsequently.

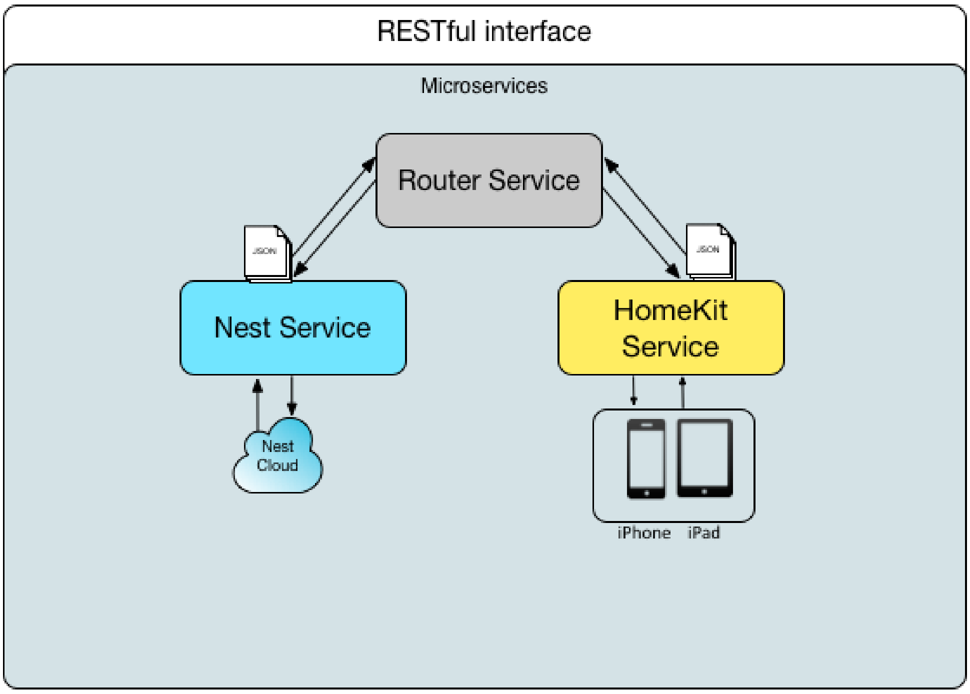
3. The integration library won't require any change or modification in the future.

However these assumptions are too risky to be considered for an always running system, because the likeliness of one of the above problems to happen is very high.
To face these problems we adopted a **microservice** architecture style to implement the integration with other ecosystems. In the case of *Nest* is not very significative but it will be much more for *HomeKit*, as we'll see later.

Microservice Architecture

Figure 3.3 illustrates the architecture proposed in our solution. All the microservices can be reached through a router service, which doesn't expose the real location of each service. This approach moves the service finding logic outside of *Calvin*, simplifying the work of the actor who won't have to deal with possible re-deployments or re-locations of the service, it will just need to know where to reach the routing service. This task will be deferred to the *routing* service, which will act as a proxy to the other micro-services.
Each ecosystem has a microservice dedicated, this is due to the different internal structure

Figure 3.3: Microservice Integration architecture



of the different technologies. For example *HomeKit* uses a hierarchical structure divided in the following, from above to the bottom: **HMHome**, **HMRoom**, **HMAccessory**, **HMService** and **HMCharacteristic**. Which differs from the simpler one of *Nest*, where there are mostly Structures and Devices.

Moreover structuring the integration with ecosystems introduces the capability to link more accounts, more devices, to the *Calvin* runtime.

The microservice API

The API to describe the connection between *Nest* and the microservice resembles the former actor structure. Following the classic *RESTful* style, to access a device will look like

```
1.
Show details about device
GET /device/<deviceid>

Example Response
{"device":
    {
    "name": "DFF9", "temperature": 37, "humidity": 50, "fan": false,
    "mode": "heat", "online": false, "serial": "fake9396122EDFF9",
    "where": "basement","structure": "HomeTest", "target": 25
    }
}

2.
Show property value of a given device
GET /device/<deviceid>/<property>

{ "value": 37 }

3.
List all the devices
GET /devices

{
   "devices": [
        "4A32","12DE","DFF9",
        "18B430AAFAA447E5","18B430AA542D0893",
        "fakebb555fb772ce405b91729b3c4ac1962b"]}
4.
Set a value for a specified property of a device
PUT /device/<deviceid>/<property>/<value>

{ "success": true }
```

3.6.4 Accessing HomeKit

HomeKit, introduced in details in the previous chapter, is the other ecosystem for which we are extending *Calvin* integration. First of all, HomeKit doesn't expose any cloud API, it's functionalities can be accessed only through their **HomeKit.framework**, a library available only for *iOS*. As can be seen the limitations compared to Nest are much more strict, which implies no direct integration. The only possible solution for the moment is to have an *HTTP* web server exposing HomeKit functionalities to the outside of the world. As previously mentioned, their library is not available on both **Mac OS X** nor the Apple TV **tvOS**, which means the only possible approach is a mobile application. However it is predicted to be supported by the former OS in the future, on which would be a more elegant and nicer solution.

The Application

Before entering in the application logic of the solution, is better to review some concepts regarding mobile development. Here are some of the architectural issue we had to deal with:

1. First of all Swift or Objective-C is not meant for building web services, therefore there are no official libraries for starting and handling HTTP requests.

2. Second, the architecture for mobile applications doesn't fit well with the standard daemon running in the background.

3. Third, the *HomeKit* accessories are accessible only from a specific component, *UIView*,which means the server has to be started in one of these views.

However on the Internet there are many open source solutions, including some HTTP servers which also supports url matching. We choose to use **GCDWebServer**[15], a lightweight HTTP server embedded in both *iOS* and *Mac OS X*.

Grand Central Dispatch

In order to fully understand how to run a web server on a mobile application we have to briefly introduce the *Grand Central Dispatch* (GCD). The *Grand Central Dispatch* is a set of tools and libraries made available from Apple for developers to handle concurrency in their applications[16]. GCD allows developers to create and dispatch objects, which will be deferred to a differed thread and queued for concurrent execution.

GCD provides three different queues on which tasks will be submitted in a form of block object; all the submitted blocks are executed on a pool of threads fully managed by the system. The three queues are:

Main : task executed serially on the main application thread;

Concurrent : tasks dequeued following a FIFO policy but executed in parallel;

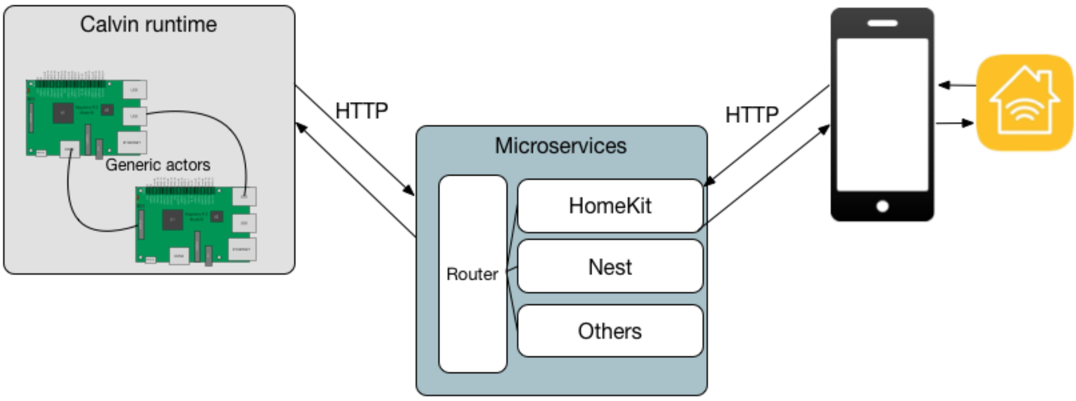
Serial : Same as before but executed in serial order;

Dispatching different parts of our system can be very helpful when we need to handle a request from a client without locking the main view of the application. In **GCDWebServer** to each handler for the different url paths, there is a block of code assigned to it, which will be handled by the GCD.

Calvin and HomeKit

The connection between a *Calvin* Actor and *HomeKit* is very similar to the former approach: the actor fires an async HTTP request, the request will travel through the whole microservice circuit and it will get the result.
The only difference lies in the number of hops to reach the device, in this case four(eight considering the returning path) including the device itself. However this implies more overhead and more latency, resulting in an overall reduction of performance, which we'll analyze more in details in the next chapter. Figure 3.4 illustrates the general architecture for *Calvin* to communicate with a *Homekit* device.

Figure 3.4: HomeKit communication architecture



Despite the overhead due to the different services, being on the same network, under the assumption of it being fast enough, the total delay should be more than acceptable.

However the architecture present a weakness,due to the single point-of-failure in the iOS application, which is the only access to the HomeKit devices.

Advantages

Integration with different ecosystems is a new topic, focusing more on the big picture rather than connecting single devices. Microservices are a bright promise for the future, for a certain category of problems and integration between ecosystems fits perfectly into it. There are some pros and cons as there are for any solution, an example may be the introduced overhead and complexity. Despite the cons the flexibility introduced my microservices is more valuable, a value for which both overhead and complexity are paid back.

Chapter 4

Validation

In the following chapter we'll discuss the results of the formerly presented implementation, highlighting the benchmarks of the solutions proposed. In the first part of the chapter we'll investigate the performance of the machine learning library, in the various situation that may happen. In the following part we will test the advantages of our microservice oriented architecture with the classical approach for integration.

4.1 Validating Speaker Recognition

This section is dedicated to the various tests to investigate the effectiveness in a home security scenario. For our tests we used two different libraries, comparing their results: *voiceid* for speaker diarization and identification, *recognito* for speaker identification only. Furthermore we have used standard configurations for all the audio files: .wav format, with a frequency 48 kHz and using 16 bit for sampling.

4.1.1 Testing the basic model

Before starting with more complex test it is mandatory to see the effectiveness of the models in the optimal situation. For the optimal situation we'll take a unique recording dividing it in the typical machine learning split: 80% learning and 20% for testing the model.
The first test will focus on a batch of voices taken from different sources:

- Youtube - A youtuber using High Quality input camera, using a model with a duration of 50 seconds.

- The integrated microphone in my laptop using a model with a duration of 16 seconds.

Table 4.1: Basic test results

	Score %	**Avg Execution (ms)**	**Avg Duration (s)**
Youtuber	92.7	3217	56
Myself	81.75	2524	9.25
Voxforge	100	2665	5.5

- Voice recordings from Voxforge using a model with a duration 32 seconds.

In this scenario the model audio durations will be of different durations and different quality, we will make more specific tests later in the document.

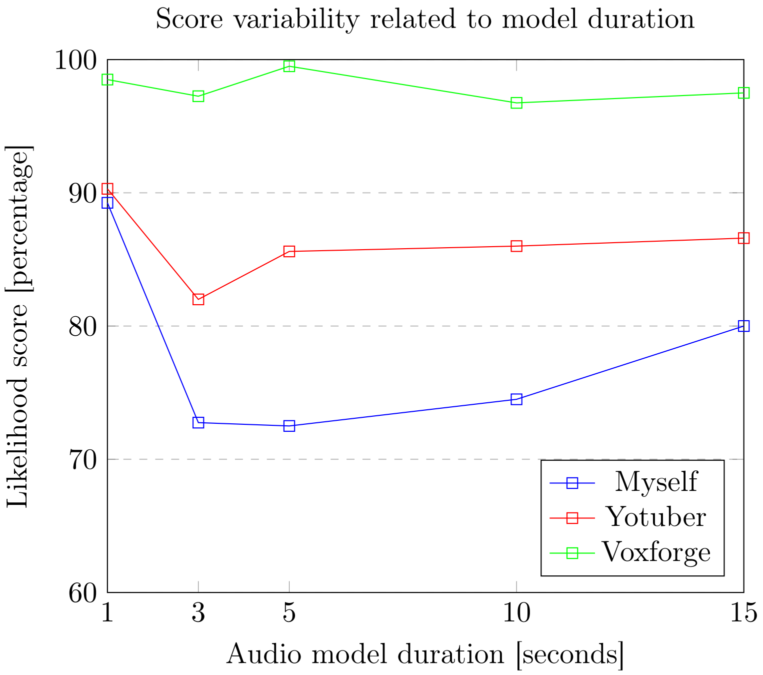
Results

In the following table we summarize the results from the test, where *score* is the likelihood percentage: less than 40% is unlikely or wrong and 100% is completely sure. The remaining fields are the average execution time for the different tests and the mean duration of the test files.

As can be seen in 4.1.1, the *recognito* library works as expected with different sources and different quality, proving its correctness for the basic functionalities. However, we can see a difference in the score of the results depending from the source of the audio recording. The less quality input, the microphone, has a lower score percentage which is very likely due to the input quality. It is indeed interesting to notice the perfect score of the audios taken from *Voxforge*. Although their quality is lower than the Youtube video, their precision is remarkably high. It is also important to notice the discrepancy between the model audio durations, which doesn't seem to be affect the score too much. We'll try to see which is the best duration for a model in the next subsection.

Finding the optimal length for a Model

The audio recordings used as model in the former chapter are all of different durations. However, in order to have a more refined comparison we need to know the best duration of a model for the best match.
The test will be executed trimming the audio model file in different slots of time, relatively: 1 second, 3 seconds, 5 seconds, 10 seconds and 15 seconds. The models are the same ones from the previous test.

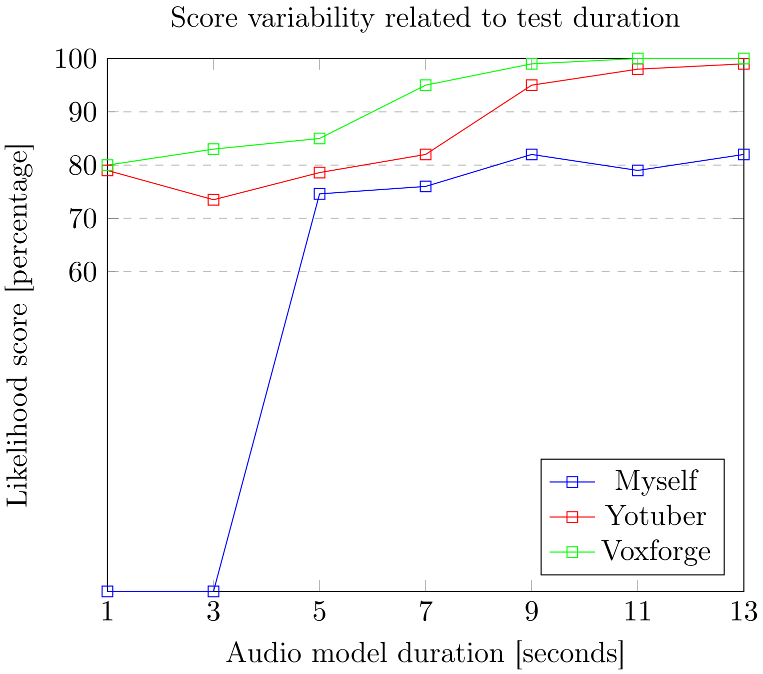


Surprisingly, as can be seen in the chart above, we had the best results with the 1 second audio model duration. However it is not a good choice to consider this length, which in some test recordings led to a mismatch in the speaker match. This was more clear when considering different people from a low quality input (the laptop's microphone), matching 2 times the wrong person. The mismatch didn't occur with the longer audio models. It is also interesting to note the fall of the likelihood score for all the models when considering a slightly longer duration for the audio. This is can be attributed to the noise introduced into the learning process by some silence in the recording, altering the identification process. The matching rate starts increasing and slowly normalizing to the optimal matching value approximately between 10 and 15 seconds.

Despite the fact that this time we used same duration audio models, the test audio segments were each one of a different duration. In the next section we'll test which is the best duration for a segment to perform a good match, using the 15 seconds model as the best for our situation.

Finding the optimal length for an Audio Segment

As introduced in the previous sections in our scenario we decided to split a single long audio recording in smaller segments. The question is now to find the optimal duration for a segment. We will use as model each relative optimal from the previous test, and apply to them various test model of the following duration length: 1 second, 3 seconds, 5 seconds, 7 seconds, 11 seconds and 13 seconds.



As can be seen from the chart, the matching increases linearly with the test duration length, approaching the best value at 13 seconds.

For short test models the system is unpredictable, though providing generally some good matches we have noticed some mismatches (marked as 0) for low level inputs. Moreover these short tests (1 or 3 seconds) seems to be heavily influenced by external noise.

It is also interesting to notice that the running time for the identification process is very slightly influenced by the length of the duration. This led us to the conclusion that the *13 seconds* audio duration is the optimal one for our scenario.

Up to now we have used models with different speakers but also from different types of sources. In the next section we will investigate how the system works when using models applied to the same speaker but from a different audio source.

4.1.2 Testing different inputs with the same speaker

As seen in the previous tests it is now clear that the input quality influences the output of the recognition. However we don't know how much the input quality affects the matching, which we will investigate in the following test. In order to get a qualitative result we set up the validation using different audio inputs, ranging in quality and with different devices. In our test we have considered the following devices:

Table 4.2: Test with different speakers

Device	**Avg Score %**
Laptop	83
Smartphone	88
Headphones	93.6

- Laptops - usually have a built-in microphone, though the quality is not high. Furthermore we used different laptops to see an average likelihood score not biased by only one device.
- Smartphones - also have an integrated microphone, usually of a higher quality since it is one of the key components of a smartphone.
- Integrated microphone - many high quality headphones have an integrated microphone, typically of better quality then the previous ones.

Results

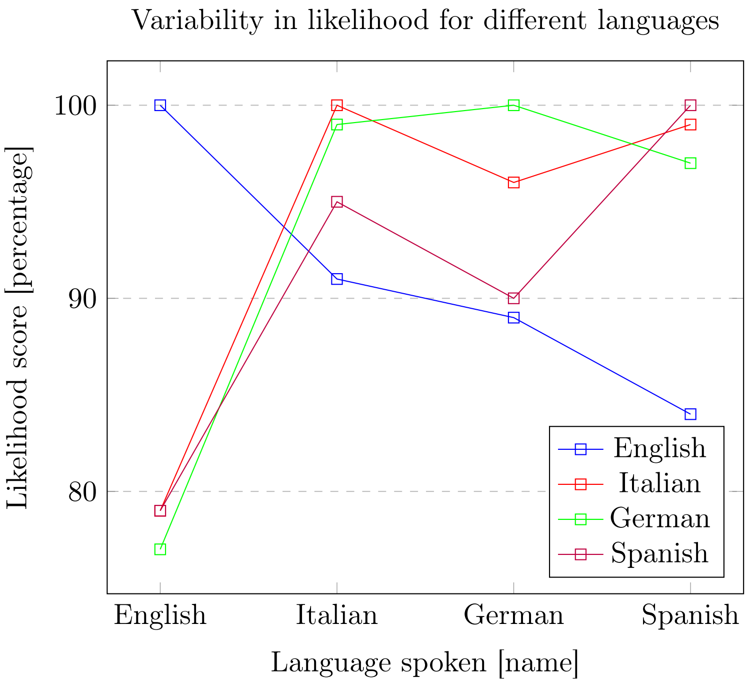
As can be see in 4.1.2 there is a definitive difference in the matching score with different inputs. The test highlighted our assumption, though the error introduced is not too high. We consider >80% already a good result, though there is only a 10% of difference between the lowest quality and the highest in the test. Therefore a halfway is more then perfect for an accurate recognition.

4.1.3 Testing same speaker talking different languages

One of the many interesting points of speaker recognition is its ability to perform recognition of people talking different languages. Although it is very unlikely for people to speak different languages in the same house, it is a characteristic to validate. In this test we will use different models, all of them based on the reading of a paragraph in different languages. Therefore we will analyze if there exists any noise or error in matching introduced by the spoken language into the model. The languages used in the test will be: English, Italian, German and Spanish. Those languages are chosen based on the speaker's ability to speak with an accent as appropriate as possible.

Results

The results showed only a small variability introduced by the language spoken. Although the speaker was using different languages the system could recognize the person without any mismatch and with high likelihood scores. It is interesting to see that the language which had the lowest match values is English. The reason may be due to the high difference in the accent with the other languages, whereas Italian and Spanish have a high relative match score.



4.1.4 Animal Recognition

Animals, and more specifically pets are one of the typical sources of noise in the house which may lead to false positives when listening for intruders. The question in this test is the ability of the system to learn from *"animals"* and recognize them as different identities. The first test to run is the basic capability of the algorithm to adapt itself to a dog's barking, and test if it can recognize the dog as familiar. Furthermore the test can be continued by testing if the model can recognize all dogs apart from human voices. We have used dogs in our examples only because it is relatively easy to find their barking recorder, though there should be no difference in using other animal sounds to perform the test (e.g. cats). Finally it is important to test the system functionalities when identifying also different kind of animals.

Animal Identification

The purpose of this test is to analyze the identification of a dog given a model and recognize the animal from it. The main purpose is to be able to identify the dog as animal, and not

Table 4.3: Animal match results

Matches	**Avg Likelihood (%)**
6 out of 13	74.7

the dog as a familiar voice. Our assumption is that given a dog barking sample we will be able to recognize all the dogs and categorize them as dog instead of human.
However the test results proved us wrong, the system can not recognize dogs from a single model, at least not precisely. From the test we had a high variance between the results, some where matching with relatively good scores (around 70% on average) meanwhile others were completely failing. Despite the non reliable result, the problem relies in the difference of dog races, specially on their size. This results in a very different and wide range of barking sounds hard to summarize in a single model. However, from the matching examples we inferred the similarity of size and race, which is a hint for the next test where we will try to verify a dog identity given its bark.

As can be seen in the table above 4.3, 6 out of 13 matched, with an accuracy of 46%, not enough for a general dog recognition but enough to open new leads for different and more accurate tests.

Animal Verification

From the experience of the previous test we can assume that there is an opportunity for the verification of an animal identity. The purpose of this validation is to test the ability of the system to recognize a familiar voice, even though it's a dog barking. This would allow us to decrease the number of false positives due to pet noise while listening for possible intrusions.
This test will use a single recording of a bark and split it using the optimal durations for model and test obtained from the previous tests. This limitation is due to the difficulty of gathering barks from the same dog in different situations, though the approach should work fine as well.

Results

In our simulation we used three different dogs barking of different sizes and races. As expected from the previous test the system managed successfully to match the dogs with their relative models. Furthermore, the accuracy of the match is very high with a minimum of 85% in the likelihood for the worst match.

As can be seen in table 4.4, the results of the different test show a high likelihood in the

Table 4.4: Animal verification

-	Avg Likelihood (%)
Dog #1	93
Dog #2	85
Dog #3	97

match. We can therefore say that speaker recognition also applies with non human voices, including pets.

Finally we can conclude that a speaker recognition system can be used to reduce the number of false positives adding the pets in the model.

4.1.5 Cheating the Speaker Recognition

Until now we have demonstrated speaker recognition working, and not, in many different situations. Another relevant question is if speaker recognition can be affected by fake impersonations, more accurately mimicking a voice. Voice imitation can be used to bypass security and surveillance systems, as in our case it would avoid the detection of the intruder.

4.1.6 Step Recognition

In recent years gait recognition gained popularity due to its non-invasivity and does not require any cooperation from the user. However, we think it's possible to apply the same thinking to steps, identifying someone by their steps. It's still a non invasive technique, and it is supposed to work similarly as it worked for the animal verification. We will start verifying the base hypothesis: is the steps are actually recognized. Subsequently we will try the identification with different pairs of shoes and see if there is any different or the steps matches anyway. Finally we will try to limit the system using different models of walking, as final verification.

4.2 Validation of the Microservice approach

In this section we will test the differences between the micro-service oriented approach with the typical integration. The microservice oriented will surely be a slower approach, mainly due to the impact of the introduced overhead. However, we will analyze the different approaches in order to establish the effective impact, evaluating its feasibility in a real world situation.

Table 4.5: Animal match results

Method	Operation	Avg Execution time (ms)
Direct	GET	0.27
Direct	PUT	0.31
Microservice	GET	4.10
Microservice	PUT	5.76

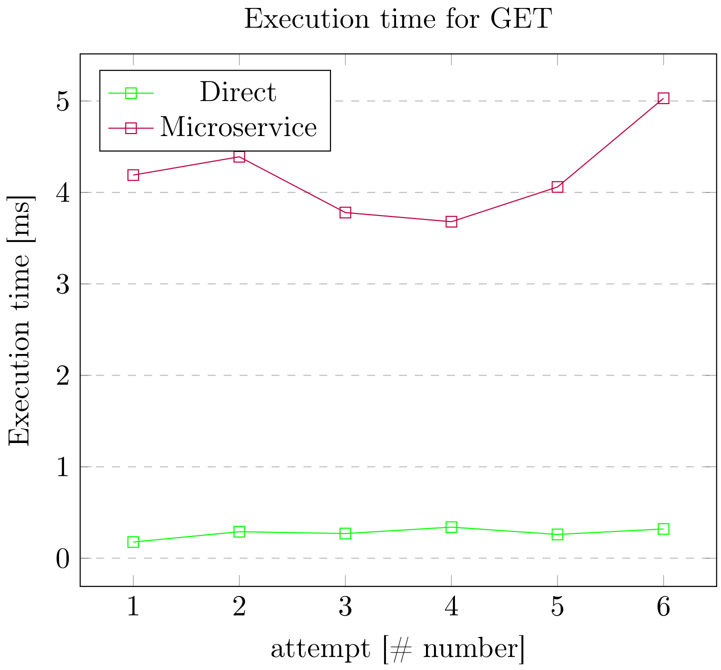
4.2.1 Microservices Oriented integration and Typical integration

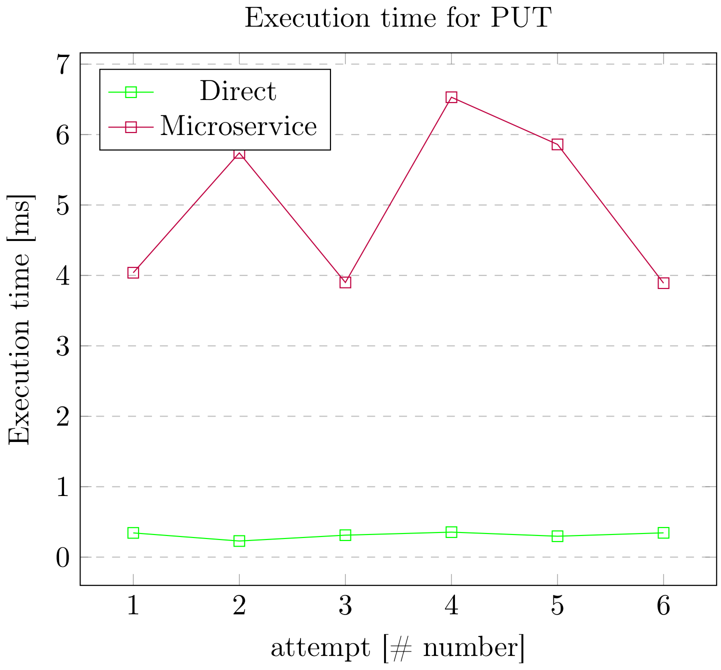
The first and most basic test to run is the running time of the different approaches, showing the real difference in the execution of the approaches. Profiling the performance of the two is the easiest way of highlighting the difference.
For this comparison we will run some reading and writing requests for some simulated devices on the *Nest* ecosystem. The test will dig into the two possible scenarios when interacting with Nest:Get and Put. More in detail:

- GET - will apply on the same device on the same property on both the tests
- PUT - will also apply to the same devices and property.

Results

As expected the running time with the typical integration is smaller. As can be seen from table 4.5 the average running for a `GET` request with the direct integration is 0.27 ms whereas the average time for the microservice approach is 4.10 ms. Similar situation for the `PUT` operation, with an average of 0.31 ms for the first approach and a higher execution time for the microservices with an average of 5.76 ms. Therefore the average delay introduced by the microservice chain is, on average, 4.64 ms, calculated on both of the operations. This is the price to pay for the different approach, a price that is worth the flexibility given in the structure. Moreover with the advent of faster networks this price given by the inter network communication will drop significantly, though even now it doesn't invalidate the overall performance.





4.2.2 Architecture limits

Our architecture, especially in the *HomeKit* situation has some limits due to the bottleneck given by the web server running on the application. Despite the fact that all the microservices can scale at need, the service dealing with the homekit devices has to run as a single instance. In this situation we will try to find the limits trying to run a high

number of requests to the end-point application and see the behavior.
In this scenario we will have many actors (i.e. from 50 to 100), all trying to read or write a casual property, stressing as much as possible the application.

Chapter 5

Related Work

In the following chapter we'll discuss the results of others who made works similar to ours.

Chapter 6

Conclusion

Just the conclusions here

This is an example for citation [AS64]

Bibliography

[AS64] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.