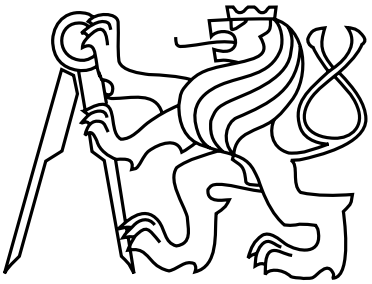
Na tomto místě bude oficiální zadání vaší práce

- Toto zadání je podepsané děkanem a vedoucím katedry,
- musíte si ho vyzvednout na studiijním oddělení,
- v jedné odevzdané práci bude originál tohoto zadání (originál zůstává po obhajobě na katedře),
- ve druhé bude na stejném místě neověřená kopie tohoto dokumentu (tato se vám vrátí po obhajobě).

Czech Technical University in Prague
Faculty of Biomedical Engineering
Department of Biomedical Technology



Master's Thesis

User Identification Method Based on Biometric Parameters of the Body

Bc. Ruben Abraham Matos Maravi

Supervisor: Ing. Anna Schlenker

Study Programme: Biomedical and Clinical Technology

Field of Study: Biomedical Engineering

May 21, 2016

Declaration

I hereby declare that I have completed this thesis independently and that I have listed all the literature and publications used.
I have no objection to usage of this work in compliance with the act §60 Zákon č. 121/2000Sb. (copyright law), and with the rights connected with the copyright act including the changes in the act.

In Kladno on May 20, 2016 .....

Aknowledgements

I would like to thank my supervisor Ing. Anna Schlenker, for giving good advises, throughout writing thesis,motivation and her support.

I also would like to thank Ing. Jakub Schlenker for his ideas and giving his time in advising during the project.

I would like to thank to my parents and family for staying loving no matter what and being supportive in everything I choose to do . Thanks to Hanna Hlushak for support and challenges we have been through.

Abstract

User Identification Method Based on Biometric Parameters of the Body

In this study we are going to investigate the feasibility of keystroke dynamics method, analysing how the language of the word (Spanish or English) could affect the authentication of native Spanish speakers, and analysing what is the impact of writing the same word in three different devices: desktop, mobile phone, and tablet device. In order to make an authentication of the participants, only two parameters were taken into account: duration and latency. We analysed the proficiency and consistency of the participants while typing words in different languages. Previous researches found these two characteristics were crucial for the authentication of individuals. The feasibility of the implementation of this method, evaluating the time taken for instructing users, error rate, and troubles occurred during the study was estimated. Our results indicated not essential difference of proficiency and consistency between English and Spanish, keeping these results among the three devices, which allowed us to obtain an appropriate result for the authentication of participants. The time consumption for the experiment, in terms of instructing participants and troubles were shown to be of a small significance.

Keywords

Biometrics, Mobile Application, Data Security, Keystroke Dynamics.

Contents

Chapter 1

Introduction

Precedent studies affirm that to have the highest degree level of security we need to combine three methods of recognition known as: something you have (Key, card), something you know (passwords or PIN) and something you are (biometrics). This theory is reflected in the Figure 1.1.

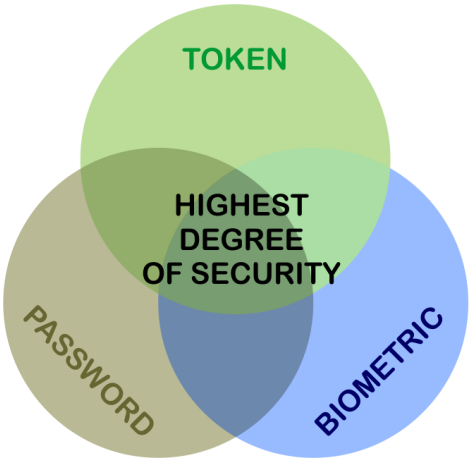


Figure 1.1: Levels of Authentication.

The general objective of this research is to find out which of those methods described above has more feasibility and reliability in terms of recognition of people in a work centre. To do this we are going to gather information from three devices, desktop, mobile phone and tablet, from a group sample and then classify them. This study will help in future for

followings researches and implementations of biometrics systems in institutes, work centres, etc.

Chapter 2

Biometric Identification Methods

For several decades researchers have been improving biometrical methods and trying to find different parameters in the human body or behavioural methods which are unique in a person in order that can be implemented in a group of people (sample) for the differentiation and identification of a singular person, getting the highest precision and a fast answer.

Basically we can say that biometrics look for data that does not change over the course of your life or that are difficult to fake or change on purpose [1]. For a biometrics to be successful need at least to have the following characteristics:

- Characteristics must not change over the course of the person's lifetime.
- Characteristics must identify the individual person uniquely.
- Characteristics need to be easily scanned or read.
- Equipment needed must not be too expensive.

For saying that a biometric system is well implemented, there are factors that have to be taken in count, for example time response, cost, precision, etc. Which depend on the hardware and as well in the software. However biometric identification is becoming commonplace as hardware and software come down in price.

In the Table 1.1 we can see some spread methods for biometric systems, the accuracy, costs for implementation of them, devices required and how good the society accepts each of these common methods.

Table 2.1: Table compares some of the biometric systems used lately, from the point of view of accuracy, cost, devices required and social acceptability. Source: http://goo.gl/4oZvda

Biometric Technology	Acuracy	Cost	Devices required	Social acceptability
ADN	High	High	Test equipment	Low
Iris recognition	High	High	Camera	Medium-low
Retinal Scan	High	High	Camera	Low
Facial recognition	Medium-low	Medium	Camera	High
Voice recognition	Medium	Medium	Microphone, telephone	High
Hand geometry	Medium-low	Low	Scanner	High
Fingerprint	High	Medium	Scanner	Medium
Signature recognition	Low	Medium	Optic pen, touch panel	High

There is a wide variety of methods for biometric identification and each of this biometric identification has its own advantage and disadvantage which are going to be discussed in this chapter.

There are two groups of biometric identification that are going to be explained (physiological and behavioural), the difference in the principle of these groups, mentioning parameters taken in count for each method [2].

Examples of both physiological and behavioural biometrics that are currently used for identification are going to be explained, giving a short overview next.

2.1 Anatomical-Physiological Characteristics

Physiological biometric is based on, as the name suggests, physiological features like fingerprints, palm prints, face, etc. The main advantage of this method over behavioural biometric is that most of these features are unchangeable so that means that over the years they will not change and they are not affected by psychological status of the person, like anger, happiness, etc. But what could affect these methods are for example injuries, accidents which can directly affect on the feature is going to be measure, for example: scratches or wounds in fingers or palms can affect the reading and recognition on the reader devices for fingerprints or palm prints.

2.1.1 Fingerprints

One of the most spread method within physiological biometry is based on fingerprints, the fingerprints recognition is the most known method thanks to its feasibility, effectiveness, reliability, fast response and quite ease to apply in a routine life.

Fingerprints are formed within the seven months of the fetus' growing and the fingerprints' boundaries configuration does not change during the life span of a human, only if the individual suffered any accident like scraping or cutting in the fingertips, also exists a probability of around 1.9×10^{-15} to find the same shape of one fingerprint.

The method for recognition of fingerprints is based on getting a two-dimensional image of your fingerprint pattern and the unique ridge patterns, a ridge is a raised portion of the epidermis on the skin, while the valleys are the white space between these ridges.

For analysing the fingerprints we need to get a clear difference between ridges a valleys in the image [3], there are different hardware technologies – optical, capacitive, ultrasound and thermal – for collecting the image of your fingerprint, from which the optical sensor is the most common used:

Optical scanner: Shines your finger with a bright light over your fingerprint, in order that the digital photography is going to be more effective and better quality, which is detected by a CCD (charge-coupled device).

Capacitive sensor: Because all fingers have ridges in the finger's surface, it implies that there are depth differences between the ridges and valleys in the fingers. In other words, the capacitive sensor creates an image by measuring distances.

Ultrasonic scanner: It consists in a transmitter and a receiver, a ultrasound pulse is emitted against the finger and some of the this pulse is absorbed while others are bounced back to the receiver, in order to measure the distance of the ridges and valleys and it is the most accurate between all fingerprints technologies and even possible to generate a 3D image of the finger.

Thermal sensor: Pyro-electric material is used for this method, material that is able to convert temperature into voltage, the principle of this method is to compare the temperature between sensor pixels which are in contact (ridges) and the ones which are not in contact (valleys).

2.1.2 Palm Prints

Unlike fingerprints, palm prints analyse as well palm creases that are present in your hand but since the pattern evaluated in palm are almost similar than fingerprint the hardware technologies for collecting the images are the same as for fingerprints – optical, capacitive, ultrasound and thermal [4]. Palm prints as well as fingerprints possess uniqueness and permanence, however palm prints have been lagged developed than fingerprints due to some restraints in computing capabilities of live-scan technologies [5].

2.1.3 Hand Geometry

Hand geometry measures and records different length, width and thickness of the hand (multiple two dimensional images) using a reading device with pegs to get the same placement of the hand each time. The main drawback of the hand geometry method is that the measurements can match with another individuals as well so that is why is not consider as a unique characteristic for this reason works combined with another form of identification as identification cards or personal identification numbers and as can be seen it is normally use for verification task rather than identification. The reading device uses a CCD camera for taking both the top surface of the hand and the side image [6].

2.1.4 Iris Recognition

Iris patterns are considered as unique pattern in human, since the probability to find a formed identical irises is on 1 in 1078, and iris patterns are formed by 10 months of age. The reader device uses a high quality digital camera and it acquires the image from the iris by illuminating it by the near infrared wavelength band to avoid harming or discomfort to the subject. Once the image is got a 2D Gabor wavelet filters and maps the segments of the iris into local amplitude and phasors. Then this Iris patterns using phase and amplitude information are described in an “Iris code” that contains a 256 to 512 byte digital template. The main drawback of the iris recognition is that Cataract surgeries change iris texture that leads to a mistakes and bad readings [7].

2.1.5 Retinal Recognition

Retinal recognition uses the unique pattern of veins beneath the eyeball. This technique involve a high quality camera and a low intensity infrared light that illuminates the eyeball in order that the retinal vasculature can be imaged because blood vessels absorb more light unlike the surrounding tissues and the patterns are digitalized and stored in a database. The drawback of this method is that the process of getting the image requires the cooperation of the subject, since the subject has to gaze into an eye-piece and keep staring a specific point in the visual field, furthermore results can be affected in terms of accuracy by diseases as cataracts or astigmatism.

2.2 Behavioural Biometric Characteristics

On the other side behavioural biometric is based on external patterns or reflections of an individual's psychology, in this field we can find: hand written signatures, voice pattern, mouse dynamics, keystroke dynamics [8]. These features, unlike physiological biometric, can be directly affected by the mood of the person, for example if a person is not in a good mood can have higher rate of error while typing or slow typing in general. That is why this method has in general less accuracy than physiological biometric due to person's mood can change sudden and due to different external factors.

2.2.1 Keystroke Dynamics

For example the keystroke dynamics recognition method has started to be used since the Second World War. It was used by military intelligence to distinguish based on the rhythm whether a morse code messaged was sent by ally or enemy, so in the beginning this was consider as a biometric solution to implement in terms of hardware.

Keystroke dynamics is a low cost biometric technology based on the hypothesis the different people have their own timing pattern while typing which can used to identify who is typing. The recorded data is then processed with different techniques such as statistical classification or neural network. This technology is becoming more popular in the same way that there are more users in computers, mobile phone and other personal devices. The disadvantages of this method and as well as in other behavioural biometrics is that it can

differ if the individual is whether angry, or sad, etc. Moreover unlike physiological biometrics, behavioural biometrics provide a lower accuracy.

For those reasons behavioural biometrics are less popular compared to physiological biometrics and furthermore it is considered as a verification method rather than identification and they work normally with the company with an ID cards, passwords or PINs for identification purposes.

2.2.2 Mouse Dynamics

Mouse dynamics likewise keystroke dynamics describes an individual's behaviour; and unlike keystroke, mouse dynamics has been begun to be studied more extensively the past three decades. Mouse dynamics started to gain more interest since the importance of user identification and verification in today's Internet-centred world is being increasing, protection of our accounts, bank movements, information from stranger hands is a higher priority. The application of this method is not only limited by, how the name suggests, computer mouse but it is also applicable to touch pads, on the other hand compared to keystroke dynamics, the information taken into account is less and a disadvantage as well as keystroke dynamics is that is difficult to mimic what you have done previously [9].

2.2.3 Voice Recognition

It is a method which consists in recognizing spoken words by the individual's vocal tract according to some acoustic patterns generated, which is an airflow that travels from the lung through different tissues to get until the mouth, from a person accordant to his/her anatomical characteristics like throat, mouth, vocal tract, jaw, larynx, etc [10]. And behavioural characteristics like voice tone, pitch or accent. There are two types of voice recognition which are:

Text dependent: Where the recognition is based in comparison of the speaker's voice against the speech sample.

Text independent: Which is more widely used because it does not need any subject cooperation and this method not only the recognition of the speaker is carried but the speech analysis as well is being evaluated.

The drawback of this method is that presents a low accuracy and an illness person such as a cold can make the recognition almost an impossible task because the subject's voice is going to be affected by the illness.

2.2.4 Signature Recognition

It is a method that consists in recognizing the subject who is writing his/her signature and this process can be carried out in two different ways depending on the type of technology.

Off-line or static: It consists of a signature scan from a document, where the subject has written the signature in a conventional paper, with a camera or a scanner for a later analysis taking into account the shape of the signature.

On-line or dynamic: It consists of an instrumented device that acquires data electronically (coordinates, pressure, inclination, azimuth, etc.) in real time and then some features are going to be extracted and trained for a later recognition.

On the other hand a disadvantage of this method is that the hardware needed could be highly cost and that some people do not have enough motor coordination to have a constant result in the writing.

Chapter 3

Security in Healthcare using Biometrics

3.1 Introduction

The term "biometric" has become worldwide known and also a method which appears in every day's life; because either nowadays in our workplace it is implemented as recognition of workers or there are applications in mobile phones for ensure access to your phone or to make some transactions like payments by internet. Along the history there has been always a need for recognizing who is sending messages, which is entering to an area of "only authorized person", etc. In the same way we are concern about the regularization of people in order to avoid stealing or accessing of personal to unauthorized areas, methods for recognizing people is growing and improving.

Biometrics can be simply define as "system of measuring physiological or behavioral characteristics in order to verify an individual's identity" http://goo.gl/w67Aoc and electronic devices like tablet, computer, mobile phones are becoming a part of our day's life, security is a term that have to be taken more serious since all the private information is in there and the demand of security is increasing because now systems based on password identification is at their limits and to get a more reliable system a biometric identification has to be included [11]. Mobile phones have become popular in every country in the world, and as mobile devices continue to evolve in terms of the capabilities and services offered, so they introduce additional demands in terms of security [12]. Distinctive biometric traits can

uniquely identify an individual [13]. For ensuring that a technique will have good results in the identification of a group needs to have a promising average error rates below the 5 % [12]. For improving these methods researches have been using different methods like artificial intelligence [14], multi-layer perceptron [15].

3.2 Keystroke Dynamics Collection

Keystroke dynamics is a method which is based on analysing how a person types but not in what the person is. The analysis is based in the rhythm patterns when typing for a posterior attempt for identification [16] [17]. Previous researches showed that keystroke dynamics can be used as a tool for identification [18] with promising results.

Unlike other biometric systems which tend to be expense at the moment of implementation, keystroke dynamics is cheap and this is because for this system someone does not need special hardware devices, what makes the implementation easier. The keystroke recognition, method which is used to identify an individual measuring the rhythm and speed of typing pattern, has basically two variables, which are shown in the Figure 3.1.

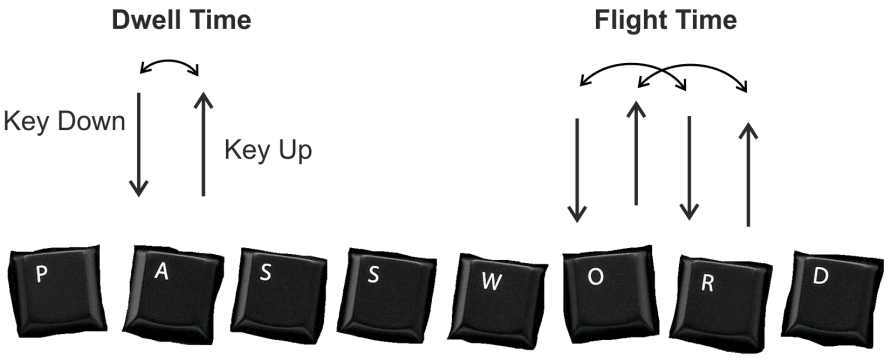


Figure 3.1: In this image we can see the Dwell Time and Fligh time are shown according to a combination of key up and key down.

- Dwell time: Duration that a key is pressed.
- Flight time: Duration between releasing and pressing the next key.

Previous researchers have identified different techniques and typing metrics that keystroke can be based. These techniques are going to be described below:

Static: This technique attempts to authenticate the user in the initial state of interaction, in other words when logging, this technique together with a password or PIN (predetermined text). Then the data acquired in the logging is compared against the previous recorded text when the person enrolled in the system.

Dynamic: This technique works different than that static, this technique attempts to still continue identifying the person whom it claimed to be even after the logging in procedure.

From this dynamic method we can describe two main type of analysis according to the duration of this:

- Continuous dynamic: This analysis extends the capturing of the data to the entire duration of the logged session [19].
- Periodic dynamic: This analysis the data during a logged session and the analysis can be constant, or part of timed supervision.

On the other hand not only duration is factor at the moment of the analysis, but the analysis can be carried out according to:

Keyword latency: When consider the overall latency of a word or a group of diagraphs and tri-graphs.

Digraph latency: When consider the latency of two successive keystrokes.

Tri-graph latency: When consider the latency of three successive keystrokes.

The analysis can be carried out analysing duration and latencies of each letter but as well sometimes taking in count another terms for the authentication process like false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (EER) which describes the efficiency of a system. These are going to be describe next and shown in the Figure 3.2.

False acceptance rate: A false acceptance event occurs when the biometric system incorrectly identifies an unauthorized person as if it was the right person. The false

acceptance ratio is the ratio of the number of times the system identified erroneously the person to the total amount of attempts that the system received. This event is considered as the most serious problem in biometric since when it happens, means that by mistake the system provides information that the person who is typing does not have the right for getting the information hence it could be dangerous and undesired for the company where the biometric system was installed.

False rejection rate: A false rejection rate event occurs when the biometric system incorrectly rejects the authorized user. The false rejection ratio is the ratio of the number of attempts the system rejected erroneously the person to the total amount of attempts that the system received. There are many reasons that why this occurred and the most common is when at the moment of enrolment of the user, the template provided was in bad quality [20].

Equal error rate: This terms means when the values for both acceptance and rejection ratio are equal.

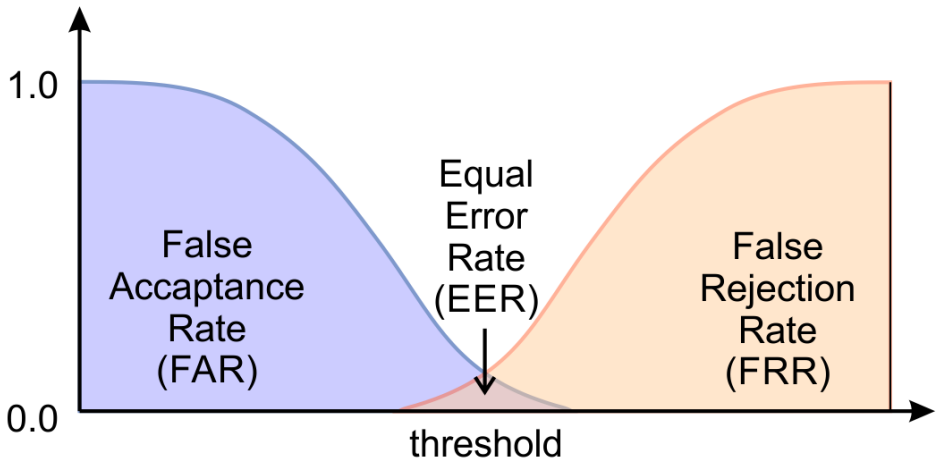


Figure 3.2: In this figure we can see the relation between the False Acceptance Rate the False Rejection Rate and the Equal Error Rate (adapted from http://goo.gl/VZ2U9z).

Performance of behavioral biometrics is measured by "False Acceptance Rate" (FAR), and False Rejection rate (FRR). But, more data can be added like: frequency of error, method of error correction, pressure(need a special keyboard), etc.

Behevioural biometrics are patterns in the way we do regular tasks and should have the next properties: Universality, distinctiveness, collectability, acceptability, transparency and minimality [21].

3.3 Benefits of Biometrics

Biometric is an important and well developed technology available for security reasons and as it was explained in the Chapter 1 for having the highest degree of security there are three fundamentals techniques [22] used for identification:

Something you know: In other words that you know something, generally passwords or Person Identification Numbers (PINs), but nobody else does.

Something you have: In other words that you possess something unique, generally tokens or different card technologies, but nobody else in your workplace or centre of studies does.

Something you are: This is what biometrics mean and there are different methods for measuring different person's characteristics like how he/she speaks, looks like, writes, etc.

Nowadays many workplaces opt for a combination of at least two techniques for a good identification of their staff, but this technologies are not only available or can be handle for adult only but also can be implemented in study centres and with children instead of lunch passes, library cards, or different places where cards are needed; a biometric technology can be applied and also avoiding different inconvenient like losing cards or breaking them [23]. But other places like libraries, sport centres also implement these techniques since they do not bother to people who are coming.

Biometric emerged from the point of security in the same way that why token and cards were developed because sometimes passwords or PINs could not bring to people enough security and in that moment cards were a good option since you always have to have it, but it was not enough and since new technology is developed also some new problems and inconvenient start appearing and this is the reason why biometric has been researched during several years but it does not mean that problem stopped appearing and this is why different

method for biometrics are being developed, looking for comfort, friendly use, easy to deal with for people and as well for a higher security of the information.

As mentioned, biometrics need to accomplish certain characteristics for a later implementation, which we can mention some of the more relevant to know if the system is appropriate or not [24].

Security: To ensure that the level of security of the biometric system reached the requirements and also resistance to possible attacks.

Privacy: To ensure that the biometric system is working according to the privacy regulations and that the personal information is ensured as well as some other related information.

Usability: To ensure that people accept the biometric system and if they are able to interact with the system without major inconvenient.

Cost/benefit: To ensure that the relation between the cost of the system corresponds with the performance.

Biometric methods were studied for the purpose of replacing the concepts of "what you know" and "what you have" for the only question of "what you are" and thanks to all those advances in this field biometric advantages like performance and reliability, have been increasing until the point that at the present time biometric is considered as an essential method to implement in companies, hospitals, universities, etc. In some cases, companies choose to have a well implemented biometric system, because the security that biometrics offers is higher than in other techniques and the cost of implementation is also decreasing.

As can be seen, biometrics technology has become a really trustworthy technology among users and also bringing them many different benefits in company with a good experience like:

Reducing fraud: When a biometric technology is implemented, this is able to detect when someone is trying to obtain a fraudulent identification and avoiding future problems with information.

Increase security: It is known that biometrics can provide a higher degree of security in comparison to other techniques and this is because to trick a biometric system someone has to gather physiological or behavioural information of the subject but even though,

nowadays biometrics system are able to detect whether is a fake information (picture of a face, voice recording, etc.) or it is the subject's real information.

Reduce inconvenient: Because it eliminates problems of holding a card every time you need to access to an area moreover posterior problems such forgetting or losing the card and also could replace the problem of hard-to-remember passwords or issues related to passwords or PINs like sharing or observed passwords by people who do not really need access to certain information.

Among other benefits: Such easy implementation of most of them because some of the methods do not need special hardware or someone who supervises the performance of the biometric device, a friendly user interfaces because for a biometric identification users do not need a lot of instructions for manipulating the device or do not have to follow a lot of steps for giving the data that they were asking for, between others.

Biometrics technology has been applied mainly to companies, corporations and different centres but it is a fact that according to how technology advances it is going to be able for every single person, bringing more secure to people personal information and more. In conclusion biometrics is an effective technology now and also with promising benefits in the future.

3.4 Biometrics and Security

Biometric is used with the idea of enhancing security in different applications that among others is the protection of personal data by limiting and monitoring access to the data using different characteristics of the human being [25] [26]. Nowadays, the use of computer networks has been spreading and at the same time the different network applications and the more the network embraces many people the more concern for identity theft problems [27].

Several methodologies for secure information or accessing exist but the major difference between those methods and biometrics is that characteristics used in biometrics are truly unique and irreplaceable, this is why biometrics is able to provide very strong access control security solution satisfying authentication, confidentiality, integrity and non-repudiation [28].

Biometric is seen as an integral part for many applications and with the possibility of improving the applications and our lives as well and there is an estimation that around 770

million biometric authentication application will be downloaded each year by 2019 only for mobile devices.

Furthermore there will be an incorporation of biometrics in mobile devices that will drive the global market with nearly 4.8 billion biometric devices by 2020 [29].

Not to mention that in order that internet will continue growing, as well as its applications, biometrics technology will be improving in the same way. Biometrics technology is still on the raise and there will be in the future more demand for this as the costs are coming down and the versatility of the systems improve, and this will lead to a better acceptance from the public in general.

3.5 Biometrics in Healthcare

Biometrics technologies have had a big impact in the healthcare field because it brings more security, convenience and also it increases the organization within the hospital, that is why the implementation of a sophisticated biometric system in a hospital is essential not only for controlling the entrance of the staff but for helping in the digitalization of all patient health records.

The digitalization in the healthcare field means the integration of different areas in a hospital for example accounting and payroll, digital X-ray database, patient records, etc. There are many advantages in the full digitalization of an enterprise, corporations or in this case hospitals, but at the same time there are some drawbacks in the implementation, the migration to a full digitalized hospital leads to a really huge exchange of information from different areas that at the same time could lead to undesired results such unauthorized access or impostors.

Patient records do not only store information of about how many times the patient has been attending to the hospital and who was the doctor that treated him/her but also patient records store information of diagnostics of the patients, X-ray images, CT images, etc. At this point with all this information future doctors know which kind of procedure has to be carried out and in case that there is going to be a surgery explain which organs are going to be involved or maybe which type of invasive method has to be done. After all we can say that not only personal information of the patient is going to be stored in the database but also vital information needed for doctors and depending on this the patient's health as well.

Electronic Health Records (EHRs) are the electronic version of patient information like previous examinations, laboratory tests, medications, etc. EHRs systems were designed for an improvement of the capturing of data across the time and there are greater advantages of an electronic information of the patient from the perspective of doctors, physicians, nurses, and hospital staff, advantages like information available whenever and wherever, reminding of meetings, ordered medications, saved information of a long term changes in the patient, etc.

The application of security systems in the healthcare is not only limited to patient records but for medical devices or patient management as well and many users will need to log in to get certain type of information. Here the problem arises since not all of them must have the same rights, for example a general practitioner will not have the same requirements as cardiologist or physician specialized [30]. It is for that reason that the healthcare system must be able to discriminate between the people who is asking for access even when the software is designed for a specialized area within the hospital.

One of the primary concerns is to maintain the data integrity to match patients with their data [31] and to show correctly to the person who is requiring the information, at the same time understanding the realities of what the patients want or expect form the digitalized system.

Furthermore the problem with medical identity theft is an issue which is occurring at this time and some of them could be related to the loss or theft of mobile devices [32] and this is where biometrics could replace old systems like passwords or PINs for authorization and accessing to specific information and avoiding excess of information to be shown to a user who does not have the right for this.

3.6 Sharing Password

Password sharing is a significant problem that is occurring today to different organizations and in hospitals it is a factor really important since all the stored information is vital for physicians specialized and that they based their criteria in what they can find in the database. Another issue within the hospital is that some passwords for different software licenses are shared as well; those problems could have a big or low impact depending on the type of information is being shared.

One of the biggest problem with a shared password system is that the proliferation of the password is a latent problem since some unauthorized people could observe the password or because a hard-to-remember passwords.

Sometimes a password can consist of a combination of letters, capital letters, numbers, special characters and because of the complexity of the password users decide to write down the password in a piece of paper, creating a gap in the security of the institution and unconsciously creating a problem of proliferation of information that undesired people could take as an opportunity to get into the confidential information and with the possibility of modifying or stealing that information for personal or someone else's benefits and conveniences.

Among different industries, normally IT leaders believe that a sharing password is not occurring within their organization. However, in the Healthcare industry about 22% share their passwords and this can occur for several reasons not only because they must have right to enter to determined software or systems but because they can delegate some simple work to their co-worker or just simple it can occur when the co-worker asks for the password.

There have been some cases where the sharing password events were a bit more than just a bad moment, cases in where a bad procedure was committed and the patient's health was compromised furthermore the patient had a bad prognostic; or when a patient's health data are shared without patient's knowledge. Moreover there is a risk of showing out important patient's information by mistake or theft.

To have an identity access management is needed to be able to differentiate between people who are entering to the information needed and at the same time be able to discriminate the type of information is going to be shown according to the rights that the user has. An implementation of identity access management will help to an improvement of the data traffic within the specific area and will provide a better experience to the user at the moment of searching the information needed.

As shown above a sharing password system can show many disadvantages in terms of privacy and security, there are some gaps in the system for showing databases with information of the patients that still have to be solved, and because creating many different passwords for sometimes the same access is a lot of password administration costs and also creating a lot of hard-to-remember password is not a viable solution. One of the best ways for identifying who is the one writing the password in that moment is to implement a type of behavioural biometrics that for this case one convenient method is keystroke dynamics

since what this method evaluates the ability of the person in writing a text and would be able to authenticate the person who is typing.

3.7 Android Programming

Android is a free software platform, it is not a hardware platform, which was created for mobile devices [33] [34] and is currently developed for Google and built on Linux-Kernel, and basically designed for touchscreen mobile devices, with an underlying operating system (OS) written in C, C++ and Java. Since the public release of Android in 2008, Android has had many updates that improved previous version, adding new features and fixing bugs; and in company with huge increases in developing of applications "apps", only in 2013 around one million of Android application were published and over a 50 billion applications were downloaded.

Because nowadays everyone has a mobile phone so Android applications are going to be developed with new features and new focuses, there are plenty opportunities for people who are actual developers as well as for future developers.

The development environment of Android uses the Java programming language with the software development kit (SDK), and the SDK is freely available to download, which is essential for building Android applications. The syntax of the development environment is as the same as Java but with some minor variations for example in class names, in Java editions there is View Class while in Android it is named as Activity Class, and in special packages as well (Android application development).

The Android structure is based in Java and eXtensible Markup Language (XML), where in general we can say that the xml files, like the AndroidManifest.xml or the folder with different resources that stores information about string variables or information about the layout of the different activities, provide the information about the project while the Java contains the logic to handle all this information provided [34].

Programming in Android also offers an Android Virtual Device (AVD) which provides an emulator that you can model according to the features needed in the program like camera, dialing pad, internal memory, between others and as well managing appearance, dimension of the screen and size of the external memory as well.

There is a strong bound between people and mobile phones because we use different applications in them everyday. In terms of security, Android applications has emerged several years ago for example now mobile devices are allowing access to different data centre services [12]. And this bound is going to become even stronger because of reasons like social purposes or because it can provide similar services like the ones you can have in a desktop computer.

In conclusion, mobile application is being diversified more and more and use of Android in biometrics is a promising technology because operations like bank transferring, trading, accessing to different centres are being done through a mobile application [35].

Chapter 4

Methods

The experiment that is going to be carried out is explained in this chapter, following steps that are going to be explained next.

The experiment will collect data from participants, in order to explain how keystrokes biometrics is a good way for authentication of individuals in different devices and how can affect to the analysis the fact that if the user is either writing in his/her native language or not.

4.1 Experiment Overview

The experiment consists in that a group of participants will write a word in English and Spanish in three different devices (personal computer, mobile phone, tablet), while a program that will be installed in all the devices will be recording all the times, in other words the program will save the information of "how long takes" for the user to write each letter. The time that are going to be recorded can be classified in two:

P-R or duration: This is the time that represents how long the user has kept pressed one key.

R-P or latency: This is the time that represents how long the user has been "seeking" from releasing one key to the next key to be pressed.

The data collected from participants then is going to be processed using statistical approaches in order to interpret the data and give a conclusion about our previous questions.

4.2 Participants

The experiment was carried out in 5 participants between the ages of 20-24 years old (4 male and 1 female).

All the participants were Spanish native speakers, who use English in their daily lives since at least six months ago. And the programs for the experiment were installed in their own personal computers and mobile phones. For the case of the experiment in the tablet device was taken only one device since not all the participants at this moment possess their own tablet device.

All the participants have almost daily use of their personal computers and mobile phones, but in the case of the tablet device the participants have an occasionally or null use of the tablet device.

4.3 Devices

For the case of the experiment in personal computers, the experiment was carried out in different personal computers of subjects, an amount of 5 (1 MAC OS and 4 Windows OS) different computers and each participant with at least 1 year of using with that personal computer.

For the case of the experiment in mobile phones, the experiment was carried out in their mobile phones, all of them with Android mobile OS, an amount of 5 different phones and each participant had at least 6 months of using that mobile device.

And for the case of the experiment in the tablet device, the experiment was carried out in an Android mobile OS as well.

4.4 P-value Approaching

This approach is used to determine how likely or unlikely of finding the observed results when a null hypothesis (H_0) is true. And shows the measurements of how extreme the observation is. The p-values is a measure of the strength of the evidence against the null hypothesis (H_0). In our case the null hypothesis (H_0) tests how similar two samples from continuous distributions or not. The results from the p-values are in a range from 0 (no chance) to 1 (absolute certainty). When the result from the p-values are near to 0 indicates that the

evidence against the null hypothesis (H_0) is higher. The p-values is used to indicate the probability and it has to be classified for a better reading of the values, for this reason the term significance level (alpha α) is used to refer to the probability of the p-values, for this experiment the significance level will be set in 5% (0.05). A guideline of the significance level is shown next with a brief description of what they mean:

When p-value $<$ 0.01: There is a very strong evidence against the null hypothesis (H_0).

When 0.01 $>$ p-value $<$ 0.05: There is a strong evidence against the null hypothesis (H_0).

When p-value $>$ 0.05: There is some weak or no evidence against the null hypothesis (H_0).

For a better understanding and reading of the p-values, the results from the p-values are going to be interpreted and shown only in two types of answers that are going to be either 1 or 0:

When p-value $<$ 0.05: It is going to be shown as 0.

When p-value $>$ 0.05: It is going to be shown as 0.

When the result from the p-values is equal to 1 indicates the rejection of the null hypothesis (H_0) occurred. When the result from the p-values is equal to 0 indicates the failure of the rejection of the null hypothesis (H_0) occurred.

4.5 P-value Calculation

For the calculation of the p-value and the answer from the either rejection or not of the null hypothesis, we are going to use the "ranksum" [36] function provided by Matlab, which calculate those values according to the different group of data in the input depending of the data of certain participant.

4.6 Data Acquisition

During the experiment, all the participants that collaborate where told about the steps to follow in order that the program developed would be able to collect the data from them.

Each participant had to write 20 times two words, one in English ("keystroke") and Spanish ("tecleado") in three different devices, the program for the devices have different appearances but with similar steps. Procedures for having a good data from different devices from participants are explained next.

4.7 Desktop

The program developed in Java was installed in every participant's personal computer with the following icon and name keyboardhook , and once the participant double click in the icon a new window is going to be open, which is shown in the Figure 4.1:

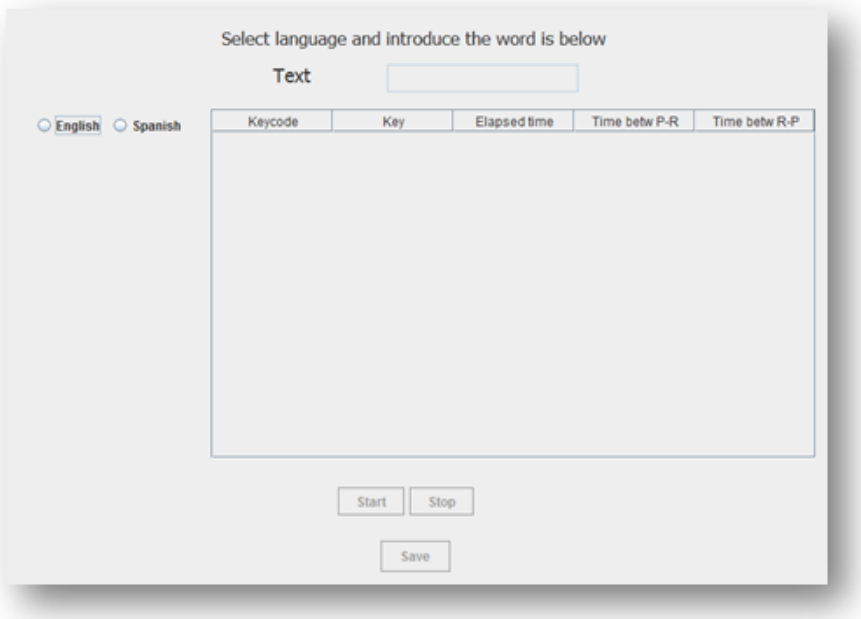


Figure 4.1: Shows the profile of the environment in which all participants write the text.

In order to start to run the program the user has to first select the language, and according to the language which is going to be selected the word "Text", that is above in the Figure

–, is going to change to either keystroke (in case the user selected English) or tecleado (in case the user selected Spanish); without selecting the language the user would not be able to start to write in the textbox or click in the “Start” button (because the “Start” button is disabled).

Once the language was selected the user is able to write in the textbox and after the user started to write in the textbox, the details of the times and key that the user is pressing are going to appear in a table, like the Figure 4.2:

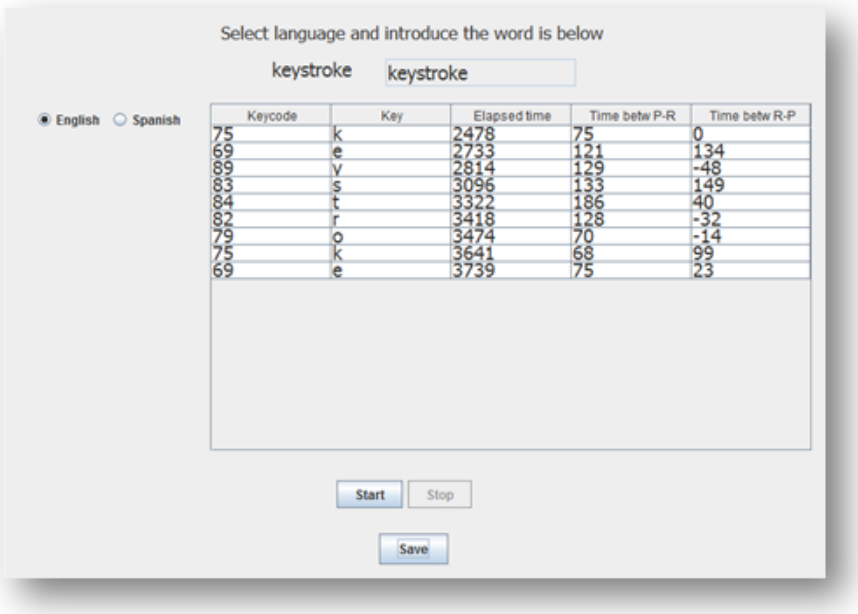


Figure 4.2: Shows the results of the times that each participant obtained after writing the rigth text.

And once the user finished writing, after pressing the “Stop” button, the user is able to save the information in the path that the user wants as long as the word that was written in the textbox matches with the predetermined by the language (“keystroke” for English and “tecleado” for Spanish).

After saving, two files are going to be generated depending on the word the user wrote. The files will have the following icon and name keystroke and tecleado , which are going to be use later.

4.8 Mobile Phone

The program for Android OS was developed in Android Studio and then it was installed in participant's mobile phones with the following icon and name 

and once the participant open the program a new screen is going to appear, which is going to look like the Figure 4.3:

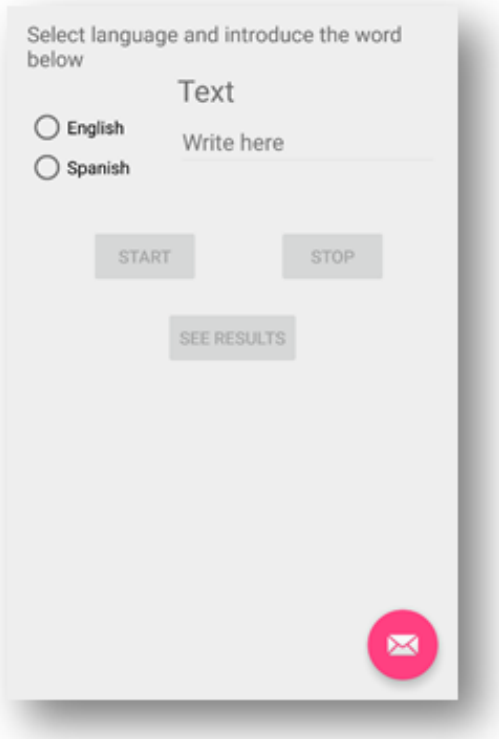


Figure 4.3: Shows the profile of the program that all participants have in their mobiles phones.

And in order to run the program, the user would need to select the language first (English or Spanish), without selecting the language the user will not be able to start writing in the textbox (where is written "Write here" in the figure) and to press the "START" button for starting to collect the times.

Once the user selected the language and pressed "START", the user will start to write in the textbox and once the user finished writing, the user has to stop the program, with the

"STOP" button, and in order to save the results the user has to click in the "SEE RESULTS" button; after clicking the next screen is going to appear.

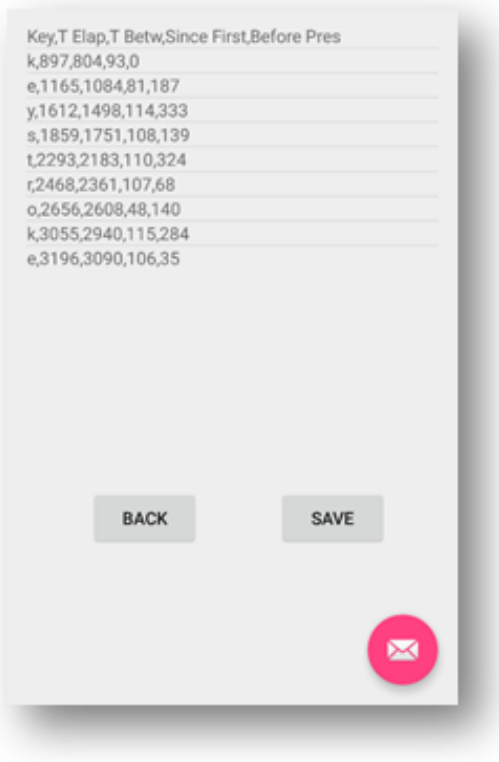


Figure 4.4: Shows the results of the times that each participant obtained after writing the rigth text.

Once the user presses the "SEE RESULTS" button, a new screen is going to show up as is shown in the Figure 4.4, and once the user is in this screen the user can press the button "SAVE" for saving the data collected, that is going to create two files tecleado.csv (if the user selected Spanish) or keystroke.csv (if the user selected English).

4.9 Tablet

Since the tablet device that was used in the experiment is Android OS as well, the environment it provides it looks similar than in mobile phone, with the same interface and the files generated after finishing writing.

The main difference between the mobile phone and the tablet is that in the tablet instead of writing when the screen is in vertical orientation, the participants were told to use in the tablet device the keyboard in horizontal orientation, basically because the lengths between keys are different than if it was in vertical orientation. This decision was made with the purpose that all the participants will experiment different lengths between keys in each device.

4.10 Data Processing

The data collected from every participant's devices was processed in Matlab and in VBA (Visual Basic for Applications), provided by Microsoft Excel.

4.11 Authentication

For the process of authentication, the time measured by the program is going to be used to enable the authentication program to make a decision if two keystrokes were made by the same user. For this purpose the VBA provided by Microsoft Excel is going to be used.

This method is based on the comparison of two numbers, if those two numbers are in the same range, so the program would understand as it was most likely typed by the "right" person. In order to have a template to compare with, the average of time difference between two letters is going to be measured. In the Table 4.1 is shown the 20 times for typing the letters "E" and "Y".

Table 4.1: This table shows the 20 times captured for the letter "E" and "Y".

E	Y
273	292
122	333
59	482
144	228
294	334
186	270
167	1096
186	356
165	313
142	227
292	632
119	250
119	250
143	274
99	481
207	228
188	271
122	249
250	248
145	250

Then the average of all the differences is going to be calculated by the next formula in Excel: "=AVERAGE("E"-"K")"

And in order to compare this average value with the one inserted by the user, the values are going to be classified according to the range that they belong to, as it is shown in the Figure 4.5.

The range that are going to be set for the classification are going to be calculated by doing iterations of the all the possible comparisons and choosing one, the one which has better results compared to the other possibilities.

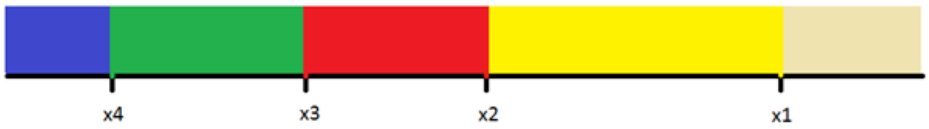


Figure 4.5: The graphic express how the ranges are going to categorize the values of the times for the participants; where x4, x3, x2 and x1 are the different limits.

Once the ranges are set, now difference between times of two consecutive letters is going to be performed, as it is shown in the Table 4.2

Table 4.2: Show the difference between the two consecutive letters, in this case Y and E, which table of times for those letters are shown in Table 4.1.

Y-E
19
211
423
84
40
84
929
170
148
85
340
131
131
131
382
21
83
127
-2
105

And if they are between the range that they meant to be so it means that it was successfully recognized or not. The process is going to be repeated for all 2 consecutive letters until it finishes in the last letter.

According to the total of matched situations for a word, the program is going to be answer of which subject is the most proximal to the template and so "authenticate" the user is currently writing.

Chapter 5

Results

The experiment was performed to compare the feasibility of keystroke dynamics for identifying the subject according to the way of typing and as well compare the feasibility between the three devices for each person. The results obtained from the devices are computed in this chapter.

5.1 Collecting Information from Devices

The three devices used for the experiment export data in the same format (*.csv). This data export a table with the parameters that are going to be taken into account for later analysis. All the devices export the same table with the same parameters.

The Table 5.1 and Table 5.2 show the next parameters:

Key code: Is the code that is assign to each key that is available in the keyboard, could be for mobile phone or desktop.

Key: It represents the key that was pressed.

Elapsed time: Is the time since the participant pressed the button "Start" till the participant presses the button "Stop".

Time between P-R: It represents the time between Pressed-Released events, in other words the time that the participant "hold" the key.

Time between R-P: It represents the time between Released-Pressed events, in other words the time of the participant "jumping" to the next key.

Table 5.1: The table shows all the parameters collected from one participant for the word "keystroke".

Key Code	Key	Elapsed time	Time between P-R	Time between R-P
75	k	1518	60	0
69	e	1787	176	93
89	y	1839	136	-84
83	s	2147	156	152
84	t	2398	148	103
82	r	2473	105	-30
79	o	2584	62	49
75	k	2757	53	120
69	e	2909	85	67

Table 5.2: The table shows all the parameters collected from one participant for the word "tecleado".

Key Code	Key	Elapsed time	Time between P-R	Time between R-P
84	t	1362	140	0
69	e	1413	110	-59
67	c	1649	111	125
76	l	1737	77	11
69	e	1887	108	42
65	a	2130	133	110
68	d	2190	117	-57
79	o	2224	59	-25

All the times expressed in the Table 5.1 and Table 5.2 are in milliseconds (ms). As it was explained in the previous chapter, for this experiment every subject wrote 20 times each word in both English (keystroke) and in Spanish (tecleado). Then the data was plotted in boxplots, in order to see how changeable the times are for the subject when writing consecutive times the same word.

The next Figures show the maximum time, the first quarter, median, third quarter and the minimum value for each letter. As well the plus symbol in red (+) shows the unique exception which is out of the normal range. In the Figure 5.1 and Figure 5.2 we can see how changeable the times are for the subject 1.

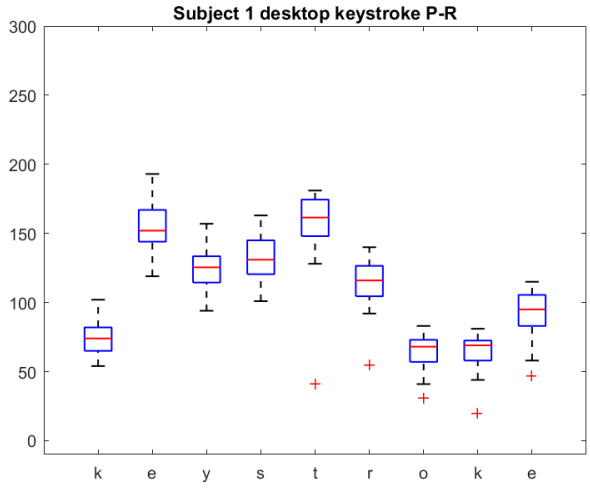


Figure 5.1: The figure shows the ranges of the times collected from the subject 1 within the 20 attempts for the word “keystroke“.

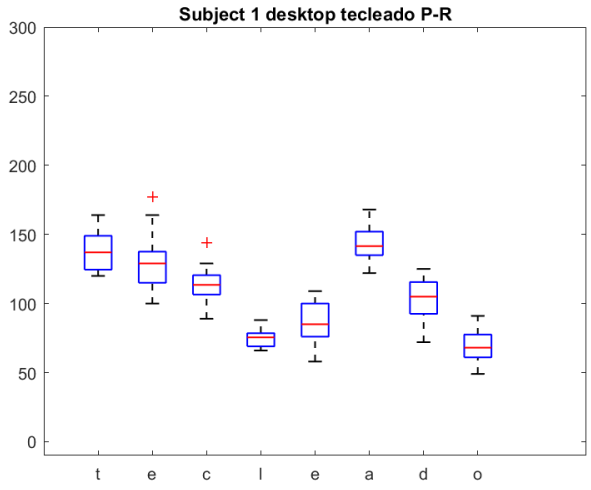


Figure 5.2: The figure shows the ranges of the times collected from the subject 1 within the 20 attempts for the word “tecleado“.

As we can see above in the graphics the range for the duration time (Pressed-Released) for each letter showed in the below part of each graphic for both words (Spanish and English).

Taken Into account the median for the subject 1, we can obtain the Figure 5.3.

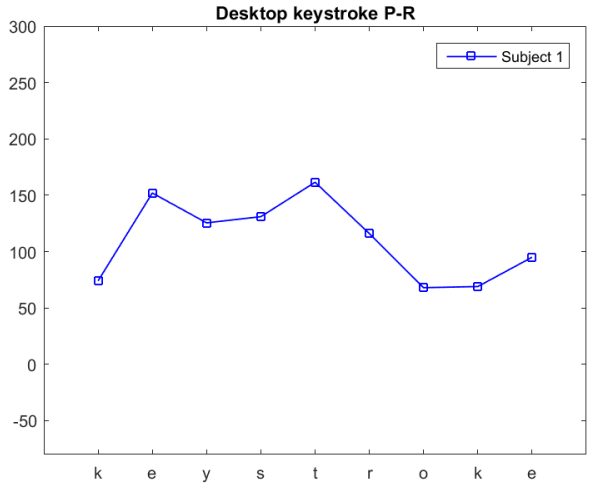


Figure 5.3: The figure shows the median of time for each letter of the word “keystroke“, calculated from the 20 different times of the Subject1.

And the Figure 5.4, there the medians for all participants.

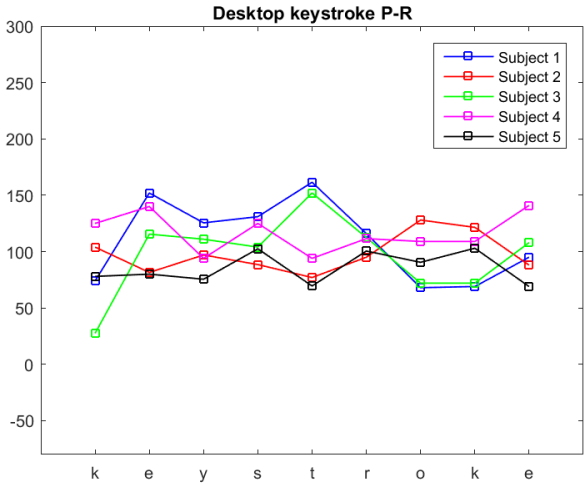


Figure 5.4: The figure shows the median of time for each letter of the word "keystroke", calculated from the 20 different times of all the participants.

In the graphics we can see that even when the times do not differ much, the trends for the each subject are different. The next step is to determine if those data could be differentiable between them.

Using the p-value function for a statistic test for determining if the trends can be likely or unlikely differentiable.

For our case the range of our p-values is the next:

p > 0.05: It is not differentiable.

p < 0.05: It is possible to differentiate.

p < 0.01: It is completely differentiable.

And according to the value that we obtain for each comparison between 2 subject for each letter of the words, the p-values can be understood simply as two interpretations, differentiable or not differentiable, applying the p-value function to the trends we get the next tables where "1" means is differentiable ($p < 0.05$ or $p < 0.01$) and "0" means that they are not ($p > 0.05$).

The analysis for each device is going to be shown next with their corresponding table of results of differentiability based on p-values.

5.2 Desktop Analysis

The results collected from different personal computers are shown in this section. Different times corresponding to each letter that is shown in the Figure 5.5, where Figure 5.5 A shows the times for the interval of Pressing-Releasing a key (P-R or duration) for the 5 subjects. Figure 5.5 B shows the times for the interval of Releasing-Pressing a key (R-P or latency) for the 5 subjects.

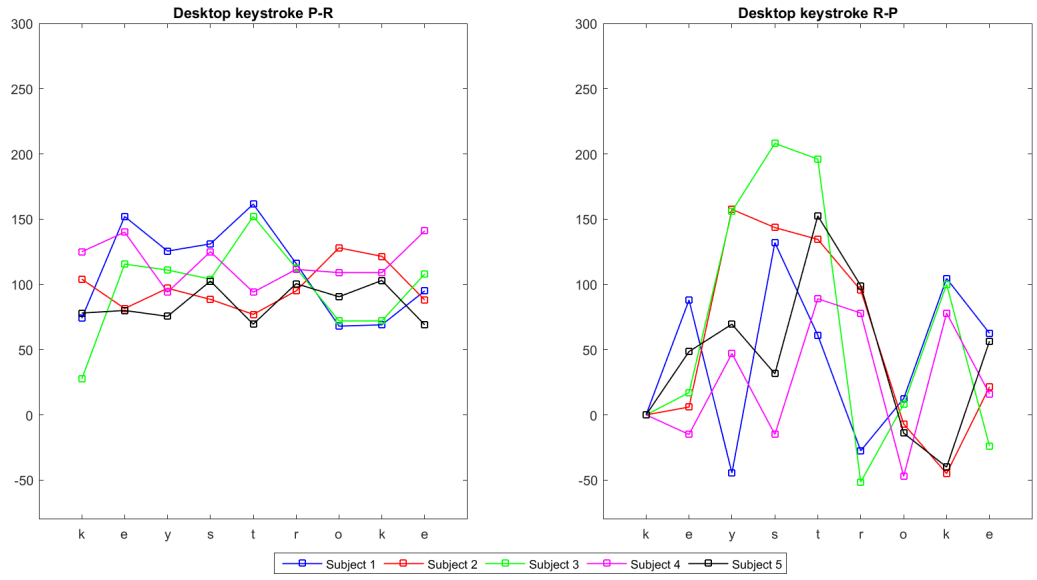


Figure 5.5: The figure shows the median of the times for all the participants. In the Figure A, it is shown the median of the times for each letter for the Pressed-Released events; and in the Figure B, it is shown the median of the times for each letter for the Released-Pressed events, both for the word "keystroke".

For illustrative purposes of the Table 5.3 we analyse the letter "K" of the word "keystroke", which trends can be seen in the Figure 5.5. In this table can be seen that the number "1" indicates that between two subjects there is a possibility to differentiate them, while the

number “0” express that the difference between the groups of data evaluated are almost similar, situation that makes difficult to differentiate them. Which in some cases can be seen directly in the graphics; like the case of the subject 1 and the subject 5 (blue line and black line respectively) for the letter “K”, which practically occupies the same spot in the graphic.

The Table 5.3 is the interpretation of the p-values function for the letter “K” of the 5 subjects.

Table 5.3: The table shows if two subjects are differentiable or not for the letter “K“, expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	1	0
Subject 2	1	x	1	1	1
Subject 3	1	1	x	1	1
Subject 4	1	1	1	x	1
Subject 5	0	1	1	1	x

The Table 5.4 is the interpretation of the p-values function for the letter “Y” of the 5 subjects.

Table 5.4: The table shows if two subjects are differentiable or not for the letter “Y“, expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	1	1
Subject 2	1	x	0	0	1
Subject 3	1	0	x	0	1
Subject 4	1	0	0	x	1
Subject 5	1	1	1	1	x

In the Table 5.3 and Table 5.4 we can see that even when in the graphics there could be seen a difference between the subject 3 and the subject 2 or 4, the table shows that the subjects 2, 3 and 4 are not likely differentiable, and this is because in the raw data of the 20 times that subjects typed, for those three subjects times were most likely the same, situation that makes unclear a differentiation between them.

The analysis for the Spanish word (tecleado) is going to be shown below.

The Figure 5.6 illustrates the trends of typing the word “tecleado” in a personal computer, where the trends for time of Pressing-Releasing (P-R) and Releasing-Pressing (R-P) are shown:

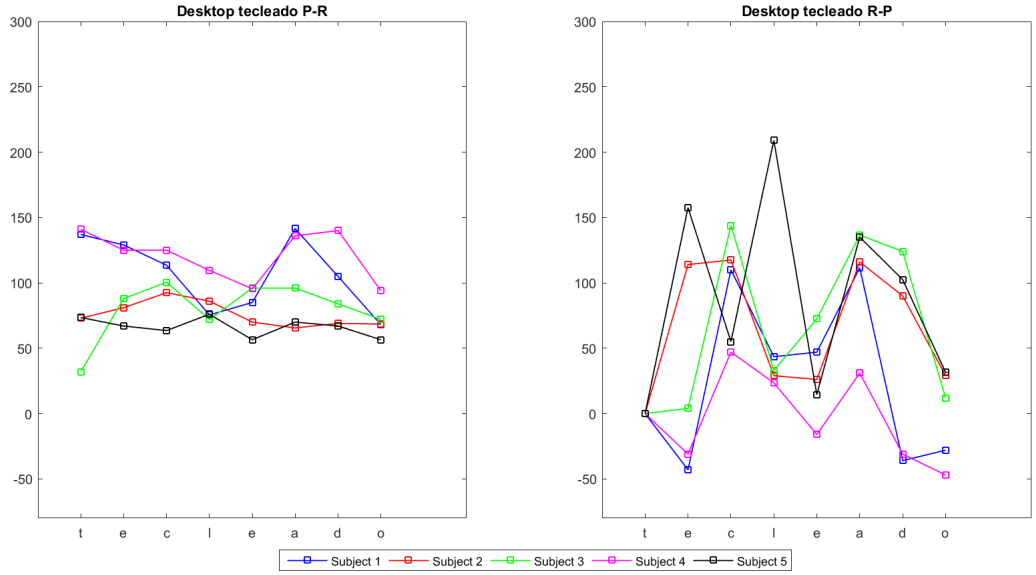


Figure 5.6: The figure shows the median of the times for all the participants. In the Figure A, it is shown the median of the times for each letter for the Pressed-Released events; and in the Figure B, it is shown the median of the times for each letter for the Released-Pressed events, both for the word “ tecleado“.

The Table 5.5 we analyse the letter “T” of the word “tecleado” and show the interpretation of the p-values from the trends shown in the Figure 5.6.

Table 5.5: The table shows if two subjects are differentiable or not for the letter "T", expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	0	1
Subject 2	1	x	1	1	0
Subject 3	1	1	x	1	1
Subject 4	0	1	1	x	1
Subject 5	1	0	1	1	x

The Table 5.6 we analyse the letter "L" of the word "tecleado" and show the interpretation of the p-values from the trends shown in the Figure 5.6.

Table 5.6: The table shows if two subjects are differentiable or not for the letter "L", expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	0	1	0
Subject 2	1	x	1	1	1
Subject 3	0	1	x	1	0
Subject 4	1	1	1	x	1
Subject 5	0	1	0	1	x

5.3 Mobile Analysis

The results collected from different mobile phones are shown in this section. Different times corresponding to each letter that is shown in the Figure 5.7, where Figure 5.7 A shows the times for the interval of Pressing-Releasing a key (P-R or duration) for the 5 subjects. Figure 5.7 B shows the times for the interval of Releasing-Pressing a key (R-P or latency) for the 5 subjects.

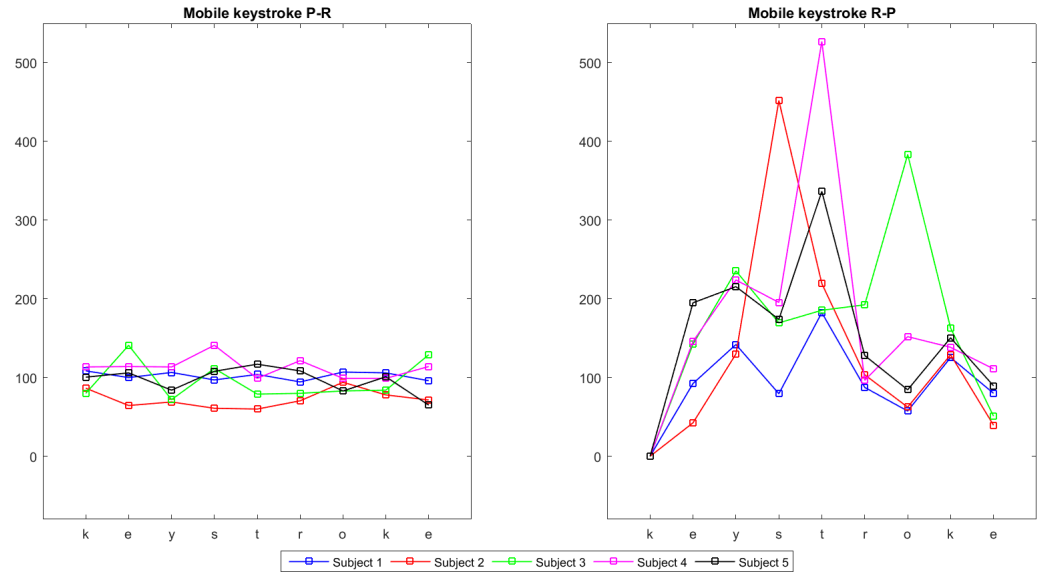


Figure 5.7: The figure shows the median of the times for all the participants. In the Figure A, it is shown the median of the times for each letter for the Pressed-Released events; and in the Figure B, it is shown the median of the times for each letter for the Released-Pressed events, both for the word "keystroke".

In the case of mobile phones the scale compared with desktops show a difference.

The letter "K" of the word "keystroke" is analysed in the Table 5.7, based in the comparison of timing between subjects in order to manifest whether those times are differentiable between each other.

Table 5.7: The table shows if two subjects are differentiable or not for the letter "K", expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	0	0
Subject 2	1	x	0	1	1
Subject 3	1	0	x	1	0
Subject 4	0	1	1	x	1
Subject 5	0	1	0	1	x

The letter “O” of the word “keystroke” is analysed in the Table 5.8 and shows whether according with the times shown in the graphics the subjects can be differentiated.

Table 5.8: The table shows if two subjects are differentiable or not for the letter “O”, expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	0	1
Subject 2	1	x	0	0	1
Subject 3	1	0	x	1	0
Subject 4	0	0	1	x	1
Subject 5	1	1	0	1	x

The analysis for the Spanish word (tecleado) is going to be shown below.

The Figure 5.8 illustrates the trends of typing the word “tecleado” in a mobile phone, where the trends for time of Pressing-Releasing (P-R) and Releasing-Pressing (R-P) are shown:

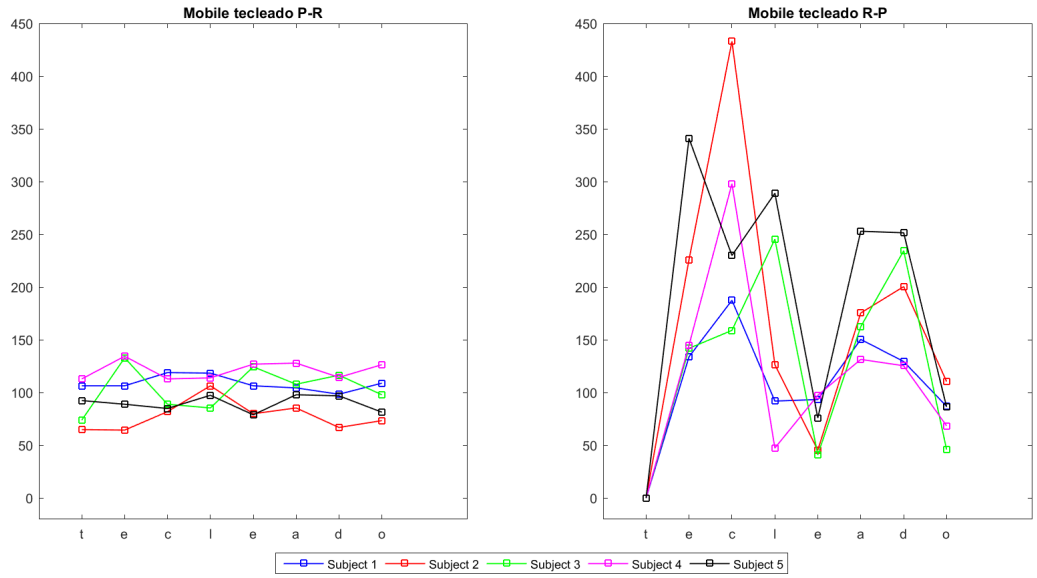


Figure 5.8: The figure shows the median of the times for all the participants. In the Figure A, it is shown the median of the times for each letter for the Pressed-Released events; and in the Figure B, it is shown the median of the times for each letter for the Released-Pressed events, both for the word " tecleado".

The analysis for the letter "T" of the word "tecleado" is shown below in Table 5.9 with their respective values for differentiation.

Table 5.9: The table shows if two subjects are differentiable or not for the letter "T", expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	0	1
Subject 2	1	x	0	1	1
Subject 3	1	0	x	1	1
Subject 4	0	1	1	x	1
Subject 5	1	1	1	1	x

The analysis for the letter "L" of the word "tecleado" is shown below in the Table 5.10 with their respective values for differentiation.

Table 5.10: The table shows if two subjects are differentiable or not for the letter “L“, expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	0	1
Subject 2	1	x	1	1	0
Subject 3	1	1	x	1	0
Subject 4	0	1	1	x	1
Subject 5	1	0	0	1	x

5.4 Tablet Analysis

The results collected from only one tablet are shown in this section. Different times corresponding to each letter that is shown in the Figure 5.9, where Figure 5.9 A shows the times for the interval of Pressing-Releasing a key (P-R or duration) for the 5 subjects. Figure 5.9 B shows the times for the interval of Releasing-Pressing a key (R-P or latency) for the 5 subjects.

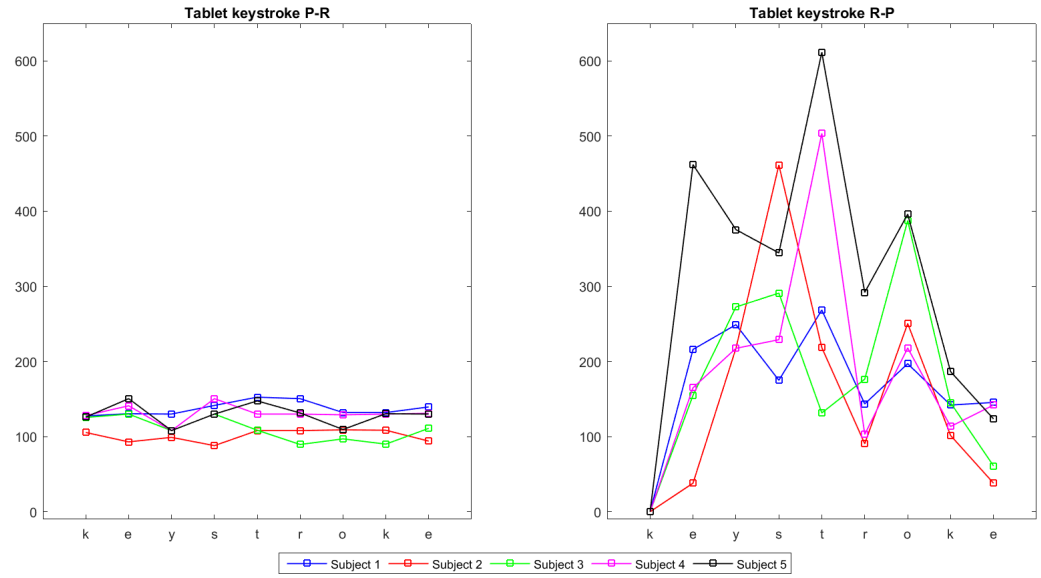


Figure 5.9: The figure shows the median of the times for all the participants. In the Figure A, it is shown the median of the times for each letter for the Pressed-Released events; and in the Figure B, it is shown the median of the times for each letter for the Released-Pressed events, both for the word "keystroke".

The analysis for the letter "T" of the word "keystroke" is shown below in Table 5.11 with their respective values for differentiation.

Table 5.11: The table shows if two subjects are differentiable or not for the letter "T", expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	0	0	0
Subject 2	1	x	1	1	1
Subject 3	0	1	x	1	0
Subject 4	0	1	1	x	0
Subject 5	0	1	0	0	x

The analysis for the letter "R" of the word "keystroke" is shown below in the Table 5.12 with their respective values for differentiation.

Table 5.12: The table shows if two subjects are differentiable or not for the letter "R", expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	1	1	0
Subject 2	1	x	0	1	1
Subject 3	1	0	x	1	1
Subject 4	1	1	1	x	0
Subject 5	0	1	1	0	x

The analysis for the Spanish word (tecleado) is going to be shown below.

The Figure 5.10 illustrates the trends of typing the word "tecleado" in a tablet, where the trends for time of Pressing-Releasing (P-R) and Releasing-Pressing (R-P) are shown:

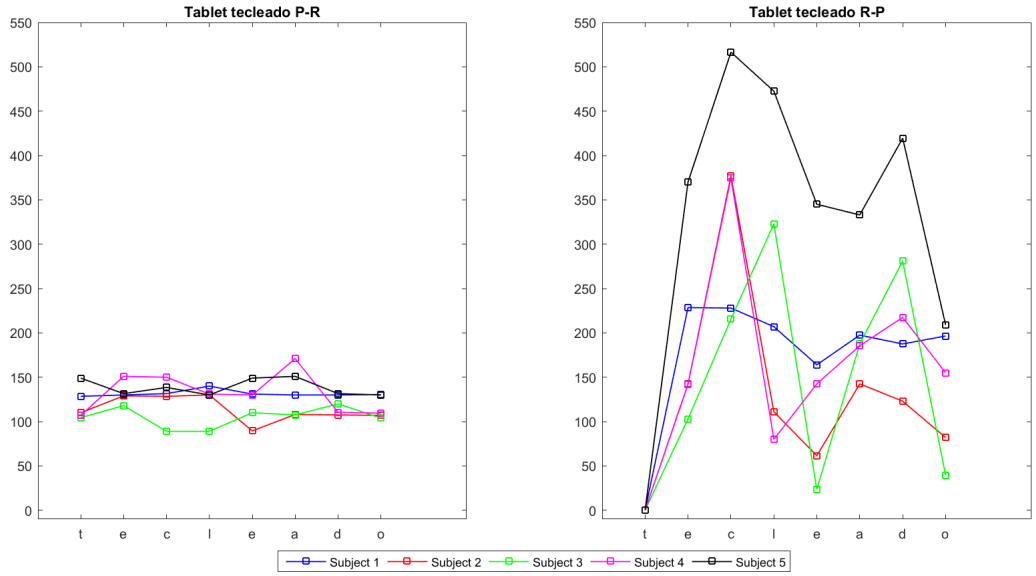


Figure 5.10: The figure shows the median of the times for all the participants. In the Figure A, it is shown the median of the times for each letter for the Pressed-Released events; and in the Figure B, it is shown the median of the times for each letter for the Released-Pressed events, both for the word "tecleado".

The analysis for the letter “T” of the word “tecleado” is shown below in the Table 5.13 with their respective values for differentiation.

Table 5.13: The table shows if two subjects are differentiable or not for the letter “T“, expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	0	1	1	1
Subject 2	0	x	1	0	1
Subject 3	1	1	x	0	1
Subject 4	1	0	0	x	1
Subject 5	1	1	1	1	x

The analysis for the letter “D” of the word “tecleado” is shown below in the Table 5.14 with their respective values for differentiation.

Table 5.14: The table shows if two subjects are differentiable or not for the letter “D“, expressed by the numbers 1 or 0.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Subject 1	x	1	0	1	0
Subject 2	1	x	1	1	1
Subject 3	0	1	x	0	1
Subject 4	1	1	0	x	1
Subject 5	0	1	1	1	x

5.5 Analysis Between Languages

As it was explained before each participant wrote two words in both English and Spanish which mean the same in the language they belong.

Each participant for the experiment typed a certain amount of keystrokes, which are indicated in the Table 5.15.

Table 5.15: The table shows how many keystrokes each participant has done in each device for the experiment.

	Spanish (tecleado)	English (keystroke)	
Desktop	160	180	340
Mobile phone	160	180	340
Tablet	180	200	380
	500	560	1060

The tables that are going to be shown below are based on the analysis of these 1060 keystrokes (in Spanish and English together) made from each participant in the three different devices.

The Table 5.16 and Table 5.17 will show the comparison of the performance of the participants for different languages when typing and for the two different times collected (duration or P-R and latency or R-P).

Table 5.16: The table shows statistical features for both words in English and Spanish for the Press-Released (duration) event for desktop.

	Spanish (tecleado)	English (keystroke)
Standard deviation	30.7657	30.7597
Average	100.7667	90.9075
Max	203	192
Min	8	1

Table 5.17: The table shows statistical features for both words in English and Spanish for the Released-Pressed (latency) event for desktop.

	Spanish (tecleado)	English (keystroke)
Standard deviation	91.3695	77.9856
Average	54.95	54.1112
Max	908	584
Min	-120	-78

To analyse how different could be, in terms of time, typing in a native language and a foreign language, we are going to analyse the standard deviation of the participant for each word. The Figure 5.11 and Figure 5.12 the correspondence of the 5 users while typing in English and Spanish

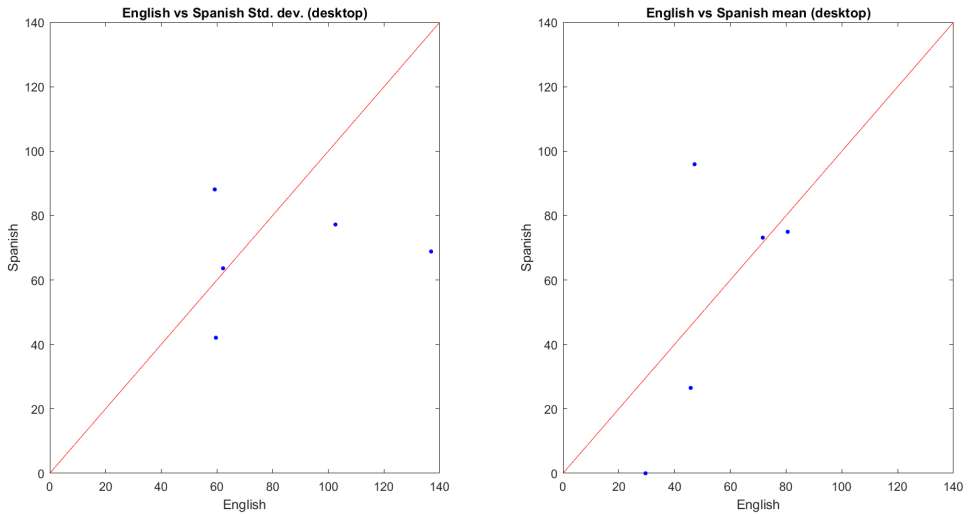


Figure 5.11: The figure shows the correspondence of mean and standard deviation for the desktop.

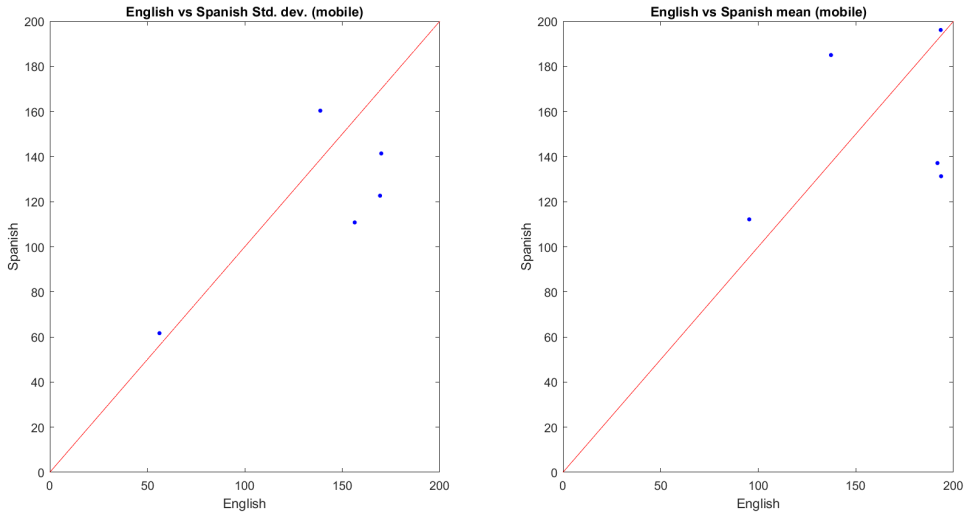


Figure 5.12: The figure shows the correspondence of mean and standard deviation for the mobile.

5.6 Analysis Between Devices

The Figure 5.13 and Figure 5.14 show how the standard deviation (Std. dev. in the figures) varies for the participants and between devices.

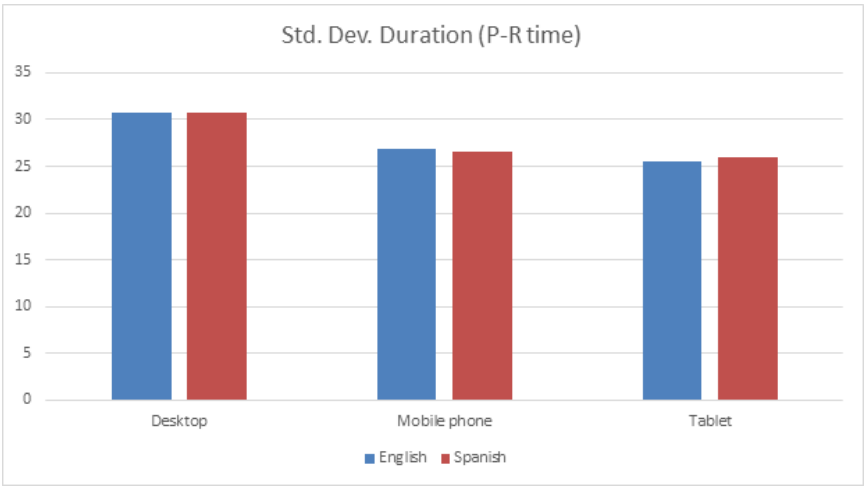


Figure 5.13: The figure shows how the standard deviation of all participant varies in each device for the duration (P-R) event.

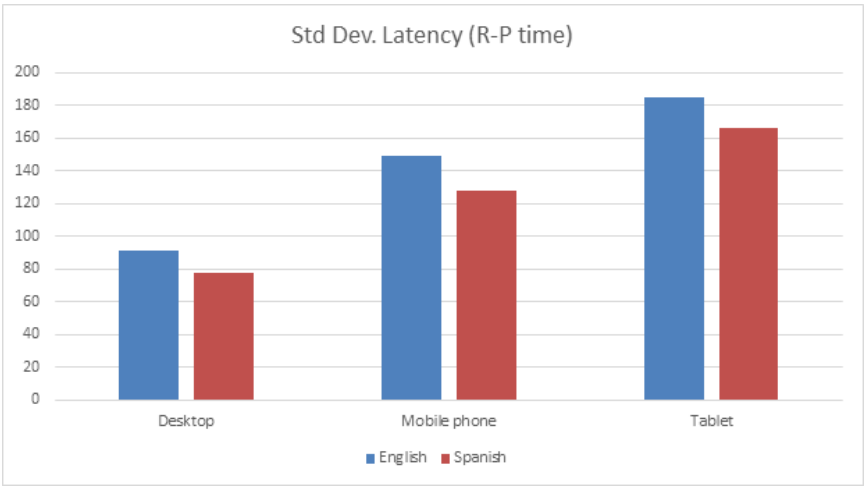


Figure 5.14: The figure shows how the standard deviation of all participant varies in each device for the latency (R-P) event.

The Figure 5.15 and Figure 5.16 show the tendency for 2 of the subjects while writing in the three devices evaluated in this experiment.

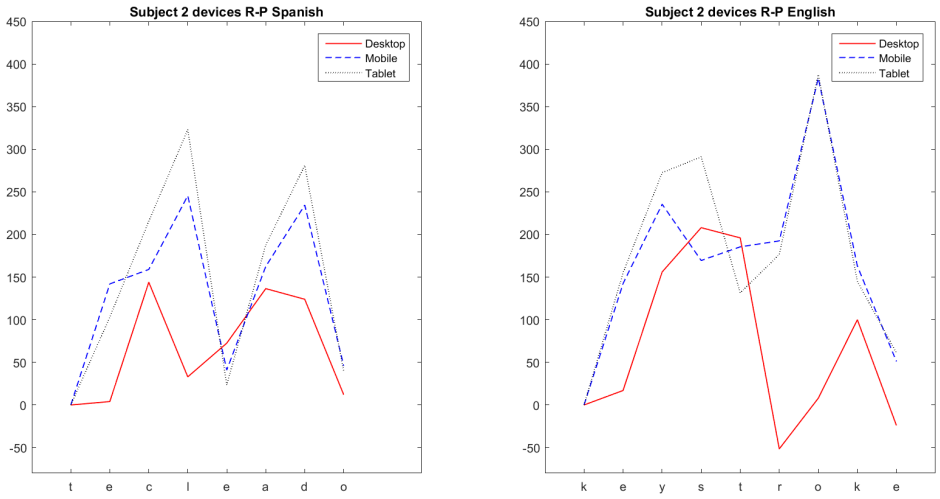


Figure 5.15: The figure shows the trends of the subject 1 while typing in the three different devices.

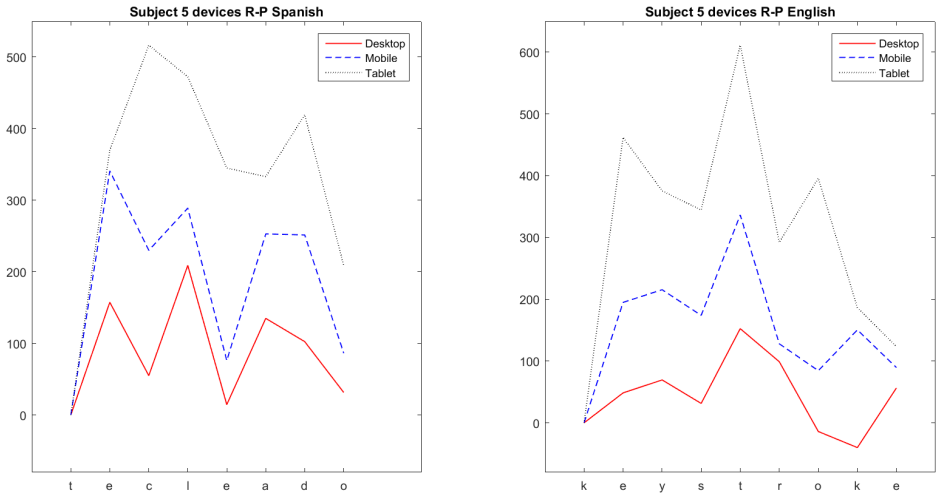


Figure 5.16: The figure shows the trends of the subject 5 while typing in the three different devices.

5.7 Authentication Method

One authentication method was implemented for the experiment in order to classify the data and give an approximation of the error for all the devices. The Table 5.18 show the result of the approximation for one word.

Table 5.18: The table shows how many letters matched in the program, in this case the subject 1 is identified as the user.

	T	E	C	L	E	A	D	O	Result
Subject 1	1	1	1	1	1	0	1	0	6
Subject 2	0	0	0	0	1	0	0	0	1
Subject 3	1	1	0	0	1	0	0	0	3
Subject 4	0	0	0	0	1	0	1	0	2
Subject 5	0	0	0	0	1	0	0	0	2

The Table 5.19 shows the limits for the different ranges and the percentage of the error rate of the program for all the devices and the words.

Table 5.19: The table shows how the error varies according to the word and devices.

Device	Word	$X_{min}(+)$	$X_{min}(-)$	$X_{mid}(+)$	$X_{mid}(-)$	$X_{max}(+)$	$X_{max}(-)$	Error
Desktop	Keystroke	18	-6	–	–	44	-68	1%
	Tecleado	10	-10	–	–	25	-25	5%
Mobile	Keystroke	29	-34	106	-114	235	-250	13%
	Tecleado	34	-34	120	-125	240	-200	15%
Tablet	Keystroke	26	-22	90	-100	230	-210	8%
	Tecleado	26	-14	105	-105	200	-200	15%

Chapter 6

Discussion

The aim of the thesis is to evaluate the feasibility of the behavioural biometrics in different devices and evaluate the performance of people when writing in their native language and in one foreign language, for this purpose, the analysis performed on the data was based on the latency of keystrokes.

6.1 Dependency on Languages

The experiment was carried out in five participants, all of them are native Spanish speakers and they wrote two words, one in English and another one in Spanish, with the same meaning.

When analysing the performance of each user when typing the two words, it was noticed in the graphics that there is no much difference of the participants when typing a word in English and Spanish, the proficiency and consistency were stable with some minor changes, and one possible answer is because the keyboard in English and Spanish are practically the same, basically the words required letters that exist in both languages and in the same position in the keyboard.

Another reason that could derive to the result we got about the proficiency and consistency when typing in different languages is that our participants have been already using in their every day's life English language since they came to Europe for at least 6 months before they took part in the experiment, hence they became more familiar with English language.

6.2 Analysis of Devices

For the five people that participated in the experiment, when analysing the standard deviation, we could see that the standard deviation of Pressed-Released events for the participants in each device was almost the same for both languages. The same situation of consistency of the standard deviation is present for the three devices but just with slight differences.

Meanwhile in the Released-Pressed events there is a clear difference between languages, hence we can say that times are more spread when a Spanish native speaker writes a text in English. And also we can see how the standard deviation is increasing according to the device that participants are writing in, that means that the consistency is decreasing for other devices different than desktop.

From Figure 5.13 we can see that the tablet device possess the highest standard deviation among the others, in other words the consistency for the tablet devices are less than mobile or desktops, and this could be because the keyboard presented in the tablet device is bigger than in the mobile phone and this can cause that the participants did spend more time in writing a word in the tablet since the keys are separated more, and comparing the keyboard of the tablet with one desktop keyboard is considerably smaller, but the difference between them in the standard deviation and in times could be explained like as following: a person feels more comfortable when writing in a desktop computer since their hands are leaning on a flat surface and they are able to use more free to use fingers when typing, unlike tablet device that participants were holding the device and their hand did not have enough freedom to move their finger across the whole screen of the tablet.

Moreover from the Figure 5.15 and Figure 5.16, we can see that even when participants did have to write in three different devices, the trends when typing a word tend to have similar deflections in each letter, which can be understood as there could be a relation between the devices that an individual will write in since the individual has almost the same tendency in writing a word, and if making a more profound analysis of this situation could show a stronger bound between all the devices where an individual could possibly write a word or a text and hence make a biometric system more robust.

6.3 Authentication

During the process of authentication, the aim was to get a system that would be able to recognize successfully the person who was writing the word, for our experiment the results were between the ranges of 0% to 15% of false acceptance rate (FAR), analysing basically the latency parameter, since the duration parameter did not have much effect when authenticating them.

And for the method implemented for the authentication of individuals, the authentication that got better results was on desktops, unlike mobile phones or tablets which got considerably lower results when authenticating the participants, higher false acceptance rate (FAR).

Even though the study did not show a clear difference about proficiency and consistency between Spanish and English, this could be explained by some reasons mentioned above, in the results of the authentication process, the authentication of the participants for the English word (keystroke) did have more accuracy than in Spanish.

6.4 Feasibility for Implementation

As it was suggested in previous chapter the implementation of the software in different devices was not an obstacle, since all the participants for the experiment did not experience any trouble when writing and the instructions that were given to the participants were easy understandable, very short and participants did not have questions after the quick explanation. Furthermore the time it took for collecting the data from the participants and from different devices was relatively short, hence participants could do it at home and without any inconvenience.

This clearly shows that the implementation is not a limitations of the keystroke dynamic biometrics, about the feasibility of the method, in the terms of costs, it is relatively cheap since there is no need to have a special hardware, and it does not generate any big problem to the user, in terms of accuracy according to the results show in previous chapter, even though I presents high number of percentage of accuracy, still a software that provides between 7% of error in average, that is still a drawback for this method.

Even though the error got in this experiment was relatively low, it has to be taken into account that the experiment counted with only 5 participants, which is a really small amount

of people compared to work or study centres. For a better performance of the method and considering a bigger sample of people, another classification method will be needed in order to get lower percentage of error.

6.5 Future Work

There are different things that could be added or improved in the experiment in order to get better results and new approaches. Listed here are some suggestions that could be taken into account in the future:

The sample of the participants were relatively low for making a more profound conclusion, and in order to get better results, the amount of participants for the next experiment should be higher.

Expand the analysis of not only two languages to compare but more, languages that will require a different keyboard than usual, in order to have a better understanding of the impact of a foreign language when typing. Furthermore another input like a longer text will be useful in order to evaluate more letters and create some vectors of two or three letters in different persons in order to have a better analysis and a better values when classifying the individuals.

For the experiment was only considered latency time for the authentication process, a usage of another variables as probably a pressure sensor in order to evaluate the force that each participant exerts in typing or another sensors, for instance a gyrometer in mobile phones and tablet, in that case the experiment will be carried out utilising special hardware available in some devices.

A more profound analysis and classification of the letters in a word or text will be needed in case of implementing the method in a bigger population.

Chapter 7

Conclusion

The experiment carried out and the results from the experiment can answer the questions presented earlier.

The difference of typing in different languages, in this case between English and Spanish did not present a big different in terms of proficiency and consistency, and this could be because the keyboard used for these two languages are similar and probably because the participants have been having continuous use of English lately. However, in future work the analysis with another language that present different keys would possible show a difference between a native language and foreign language when typing in terms of proficiency and consistency.

Meanwhile the difference of typing between different devices show a clear difference, and this is probably the different platforms provided by the devices. On the other hand in some cases there is a clear tendency for the participants between the devices, which means that for future work there could be a stronger relation between the devices that will help at the moment of implementing an authentication method using a combination of different devices.

The feasibility and the implementation of this method is a really promising technology and there could be taken into account other parameters in the future, in this study the error rate at the moment of authentication was low, and that new implementations of this biometric method in the future with more robust systems of identification is a really good option.

References

[1] SANS Institute and Kyle Cherry. Biometrics: An in depth examination. *SANS Institute*, 2003.

[2] *An Introduction to Biometric Authentication Systems.* Springer, 2005.

[3] National Sciece and Technology Council (NSTC). Fingerprint recognition. pages 100–109.

[4] National Sciece and Technology Council (NSTC). Palm print recognition. pages 121–127.

[5] IEEE, Jianjiang Feng, Anil K. Jain, and fellow. Latent palmprint matching. *IEEE Computer Society*, 31:114–118, 2009.

[6] National Sciece and Technology Council (NSTC). Hand geometry. pages 1–7, 2006.

[7] National Sciece and Technology Council (NSTC). Iris recognition. pages 114–118.

[8] Roman V. Yampolskiy and Venu Govindaraju. Behavioural biometrics: a survey and classification. Technical report, University at Buffalo, 2008.

[9] Nan Zheng, Aaron Paloski, and Haining Wang. An efficient user verification system via mouse movements. Technical report, The College of William and Mary, 2011.

[10] National Sciece and Technology Council (NSTC). Speaker recognition. pages 128–133.

[11] Milan Adámek, Miroslav Matýsek, and Petr Neumann. Security of biometric systems. *Procedia Engineering*, pages 169–176, 2015.

[12] N. L. Clarke and S. M. Furnell. Advanced user authentication for mobile devices. *Computers and Security*, 26(2):109–119, 2007.

[13] P. H. Griffin. Biometric knowledge extraction for multi-factor authentication and key exchange. *Procedia Computer Science*, 61:66–71, 2015.

[14] B. Purgason and D. Hibler. Security through behavioral biometrics and artificial intelligence. *Procedia Computer Science*, 12:398–403, 2012.

[15] S. Seob Hwang, S. Cho, and S. Park. Keystroke dynamics-based authentication for mobile devices. *Computers and Security*, 28(1-2):85–93, 2009.

[16] Fabian Monrose and Aviel D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, pages 351–359, 2000.

[17] Pin Shen Teh, Andrew Beng Jin Teoh, and Shigang Yue. A survey of keystroke dynamics biometrics. *The ScientificWorld Journal*, pages 351–359, 2013.

[18] D. Shanmugapriya and Dr. G. Padmavathi. A survey of biometric keystroke dynamics: Approaches, security and challenges. *International Journal of Computer Science and Information Security (IJCSIS)*, 5:351–359, 2009.

[19] Hataichanok Saevanee, Nathan Clarke, Steven Furnell, and Valerio Biscione. Continuous user authentication using multi-modal biometrics. *computers and security*, 2015.

[20] Henning Gravnås. User's trust in biometric authentication systems., 2005.

[21] H. Crawford, K. Renaud, and T. Storer. A framework for continuous, transparent mobile device authentication. *Computers and Security*, 39(PART B):127–136, 2013.

[22] Ing. Martin Drahanský. Biometri security systems fingerprint recognition technology., 2005.

[23] Schools learn about the benefits of biometrics. *Biometric Technology Today*, 2001.

[24] Ma. Belén Fernández Saavedra. *Evaluation Methodologies for Security Testing of Biometric Systems beyond Technological Evaluation.* PhD thesis, Universidad Carlos III De Madrid, 2013.

[25] Vassiliki Andronikou, Dionysios S. Demetis, and Theodora Varvarigou. Biometric implementations and the implications for security and privacy. Technical report, National Technical University of Athens and London School of Economics and Political Science.

[26] Pin Shen Teh, Andrew Beng Jin Teoh, Connie Tee, and Thian Song Ong. Keystroke dynamics in password authentication enhancement. *Expert Systems with Applications*, 2010.

[27] Umut Uludag. *Secure Bimetric Systems.* PhD thesis, Michigan State University, 2006.

[28] SANS Institute and Wayne Penny. Biometrics: A double edged sword - security and privacy. *SANS Institute*, 2002.

[29] Bob Violino. Biometric security is on the rise. *CSO*, 2015.

[30] Quintin Armour and Didier Thizy. Developing successful healthcare software: 10 critical lessons. *macademian.*

[31] Darrell Shawl. Biometrics - implementing into the healthcare industry increases the security for the doctors, nurses, and patients. Master's thesis, DAVENPORT UNIVERSITY, 2013.

[32] Steve Gold. Healthcare biometrics:solving the staff and patient security governance challenge. *Biometric Technology Today*, 2013.

[33] *Unlocking Android: A Developer's guide.* Manning Publications Co., 2009.

[34] Benny Skogberg. Android application development: A guide for the intermediate developer. Master's thesis, Malmö University, 2010.

[35] Hataichanok Saevanee, Nathan L. Clarke, and Steven M. Furnell. Multi-modal behavioural biometric authentication for mobile devices. *International Federation for Information Processing (IFIP)*, pages 465–474, 2012.

[36] Statistics: 2.2 the wilcoxon signed rank sum test.

List of Figures

List of Tables