



Universidad de Valladolid

Escuela de Ingeniería Informática

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática
Mención en Computación

**Investigación y Desarrollo en
Técnicas de Reconocimiento Biométrico
mediante Dispositivos Ponibles
(wearables)**

Autor:

Dña. Irene Salvador Ortega



Universidad de Valladolid

Escuela de Ingeniería Informática

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática
Mención en Computación

**Investigación y Desarrollo en
Técnicas de Reconocimiento Biométrico
mediante Dispositivos Ponibles
(wearables)**

Autor:

Dña. Irene Salvador Ortega

Tutores:

D. Carlos Vivaracho Pascual

Dña. María Aránzazu Simón Hurtado

Agradecimientos

Me gustaría dedicar unas palabras de agradecimiento a todas las personas que me han acompañado en mi camino por la Universidad, durante estos cinco años, desde profesores que me han permitido aprender lo que hoy sé y han hecho que vaya a recordar estos años como una etapa bonita y positiva, hasta familiares y amigos que han conseguido que no me sintiera sola en ningún momento.

A mis tutores, Carlos Vivaracho y M^a Aránzazu Simón, por vuestra ayuda y consejos. Me habéis guiado y apoyado en todo momento, enseñándome el mundo de la investigación desde dentro como algo difícil pero muy bonito y gratificante. A Miguel Alejandro Fernández, que ha estado a mi lado, dándome consejos desde el principio hasta el final. Hemos formado un gran equipo juntos y me habéis hecho crecer tanto personal como académicamente. Muchas gracias, sin dudarlo, os volvería a elegir.

A todos los profesores que han prestado su ayuda en este proyecto, Félix Prieto en toda la parte de Android, Pablo de la Fuente con la planificación y Luis Ángel García con sus consejos y documentación, gracias.

A mi familia. A mis padres, por apoyarme en cada decisión y hacerla posible. A mi hermano por permanecer a mi lado y ayudarme en los momentos difíciles.

A mis compañeros de Universidad, en especial al equipo que nunca olvidaré, Raúl, Adrián, me habéis enseñado lo que significa ser compañeros, aportando cada uno de nosotros esa chispa tan mágica a cada tarde haciendo trabajos. Espero que sigamos juntos siempre, celebrando cada acontecimiento que ocurra en nuestras vidas, sea bueno o malo.

A mis abuelos que, aunque ya no están aquí, no los he olvidado nunca y sé que me guían e intentan que cada decisión, me salga de la mejor manera posible, proporcionándome con cada una de ellas una lección de vida, sea positiva o negativa.

Muchas gracias a todos, sin duda, este trabajo va por vosotros.

Resumen

El creciente uso de sistemas biométricos para el reconocimiento de las personas supone un constante aumento del esfuerzo investigador en la mejora de los existentes y en la propuesta de nuevos y originales enfoques, que puedan mejorar o complementar los existentes.

En el presente proyecto se va a trabajar con una biometría basada en las características del comportamiento del ser humano que permite verificación no intrusiva, continua, fácil de conseguir y difícil de robar o falsificar. El objetivo principal es determinar si el uso de los sensores presentes en los dispositivos ponibles puede permitir o no la verificación biométrica de personas mediante su forma de caminar.

Ya existen trabajos en el tema que serán usados como referencia, pero ninguno, hasta donde llega el conocimiento del grupo de investigación, usando dispositivos comerciales.

Este trabajo se ha realizado en conjunto con otro del Grado en Estadística, siendo ambos una continuación de trabajos previos ya realizados por el grupo de investigación, en los que se desarrolló el sistema móvil de captura de datos y se obtuvo un corpus con el que trabajar. En el proyecto del Grado en Estadística se ha hecho un análisis profundo de dichos datos: extrayendo características, a partir de la prueba de diferentes técnicas de preprocesamiento, ajustando los valores de diversos parámetros y el número de características a utilizar, dejando para este proyecto el análisis del rendimiento final y la propuesta de posibles combinaciones que beneficiarían a los resultados. De manera que estos dos trabajos permitirán tener unas bases sólidas en las que asentar futuras investigaciones, proponiendo como resultado final, un sistema de reconocimiento, que servirá como aproximación inicial de lo que se puede llegar a conseguir en esta biometría.

Palabras claves

Análisis de Fourier, base de datos, Biometría, dispositivos comerciales, dispositivos ponibles, dominio de la frecuencia, dominio del tiempo, forma de andar, reconocimiento biométrico, SCRUM

Abstract

The growing use of biometric systems for the recognition of people supposes a constant increase of the research effort in the improvement of existing ones and in the proposal of new and original approaches that can improve or complement existing ones.

In this project we will work with a biometrics based on the characteristics of human behavior that allows an unobtrusive, continuous, easy to obtain and difficult to steal or falsify verification. The main objective is to determine if the use of the sensors present in the wearable devices can allow or not the biometric verification of people by their gait.

There are already studies on the subject that will be used as a reference, but none, as far as the knowledge of the research group goes, using commercial devices.

This work has been done in conjunction with another of the Degree in Statistics, both being a continuation of previous works already done by the research group, in which the mobile data capture system was developed and a corpus was obtained to work with. In the project of the Degree in Statistics an in-depth analysis of these data has been made: extracting characteristics, from the test of different preprocessing techniques, adjusting the values of various parameters and the number of characteristics to be used, leaving for this project the analysis of the final performance and the proposal of possible combinations that would benefit the results. So these two works will provide a solid basis on which to base future research, proposing as a final result, a recognition system that will serve as an initial approximation of what can be achieved in this biometrics.

Keywords

Fourier Analysis, database, biometrics, commercial devices, wearable devices, frequency domain, time domain, gait, biometric recognition, SCRUM

Índice general

Resumen	5
1. Introducción	17
1.1. Biometría	19
1.2. Motivación	22
1.3. Objetivos	23
1.3.1. Objetivo general	23
1.3.2. Objetivos específicos	23
1.4. Estructura de la obra	23
2. Plan de desarrollo y seguimiento	25
2.1. Modelo de proceso	25
2.2. Restricciones	26
2.3. Gestión de riesgos	27
2.4. Planificación	30
2.5. Planificación de las iteraciones y seguimiento	37
2.5.1. Iteración 1	37
2.5.2. Iteración 2	40
2.5.3. Iteración 3	42
2.5.4. Iteración 4	43
2.5.5. Iteración 5	43
2.5.6. Iteración 6	44
2.5.7. Iteración 7	45
2.5.8. Iteración 8	46
2.5.9. Iteración 9	47

2.5.10. Iteración 10	49
2.5.11. Iteración 11	50
2.6. Proceso técnico completo	51
3. Estado del Arte	55
4. Datos y análisis	63
4.1. Base de datos	63
4.2. Limpieza de los datos	66
4.2.1. Pulsera Microsoft, acelerómetro (ACC):	67
4.2.2. Pulsera Microsoft, giroscopio (GYR):	71
4.2.3. Reloj Motorola, acelerómetro (ACC):	71
4.2.4. Reloj Motorola, giroscopio (GYR):	72
5. Configuración experimental	75
5.1. Extracción de características	76
5.1.1. Dominio del tiempo	76
5.1.2. Dominio de la frecuencia	78
5.1.3. Señal combinada	80
5.2. Medición del error	81
5.3. Experimentos	83
5.4. Clasificación	85
6. Experimentos: Resultados finales	93
6.1. Dominio del tiempo	97
6.2. Dominio de la frecuencia	104
6.3. Comparación entre dominios	112
7. Adquisición de nuevos datos	115
7.1. Puesta en marcha de los dispositivos y la APK	115
8. Conclusiones y trabajo futuro	121
8.1. Conclusiones	121
8.2. Líneas de trabajo futuro	122

<i>ÍNDICE GENERAL</i>	11
Acrónimos y abreviaturas	125
Índice alfabético	127
Anexos	128
A. Contenido del CD	131
Referencias	133

Índice de figuras

1.1. Fases involucradas en el sistema de verificación biométrica.	19
2.1. Calendario de desarrollo del proyecto propuesto inicialmente.	36
2.2. Diagrama de Gantt con la planificación completa.	37
2.3. Subdiagrama de Gantt con la planificación del dispositivo Microsoft acelerómetro.	38
2.4. Subdiagrama de Gantt con la planificación del dispositivo Microsoft giróscopo.	39
2.5. Subdiagrama de Gantt con la planificación del dispositivo Motorola acelerómetro.	39
2.6. Subdiagrama de Gantt con la planificación del dispositivo Motorola giróscopo.	40
2.7. Uso de los recursos en la planificación completa.	41
2.8. Calendario real del desarrollo del proyecto.	52
3.1. Esquema de un <i>ciclo de marcha</i>	57
4.1. Tipos posibles de series de datos.	63
4.2. Dispositivos disponibles.	65
4.3. Formato de los datos que se van a utilizar.	66
4.4. Usuario 13, sesión 2, toma 2, coordenadas X, Y, Z respectivamente.	67
4.5. Mostrar error de retroceso en los datos.	67
4.6. Error en el tiempo relativo del usuario 19.	68
4.7. Usuario 19, sesión 2, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC)	69
4.8. Usuario 19 modificado, sesión 2, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC).	69
4.9. Usuario 20, sesión 1, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC)	69
4.10. Usuario 20 modificado, sesión 1, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC).	69
4.11. Distancias usuarios (componente X).	70

4.12. Usuario 19, sesión 2, toma 1, coordenadas Z (Micro GYR).	71
4.13. Usuario 20, sesión 1, toma 1, coordenadas Z (Micro GYR).	72
4.14. Señal usuario 4, componente Y, Motorola Acelerómetro.	73
5.1. Ventanas de suavizado aplicadas a la transformada de Fourier.	79
5.2. Aplicación de Fourier con y sin ventanas al usuario1, S1, M1, 1 ^a ventana	80
5.3. Ejemplo de EER a partir de la curva ROC y el AUC.	82
5.4. EER variando k en KNN, para cada usuario, Dominio del Tiempo XYZ.	89
5.5. EER variando k en KNN, para cada usuario, Dominio del Tiempo Módulo.	90
5.6. EER variando k en KNN, para cada usuario, Dominio de la Frecuencia XYZ.	90
5.7. EER variando k en KNN, para cada usuario, Dominio de la Frecuencia Módulo.	91
6.1. Comparación de la forma de la serie temporal en los dos sensores.	94
6.2. Esquema tablas del Rendimiento Final	95
6.3. Resultados EER medio - Dominio del Tiempo (Dispositivo MICRO).	98
6.4. Resultados EER medio - Dominio del Tiempo (Dispositivo MOTO).	99
6.5. Resultados EER medio - Dominio de la Frecuencia (Dispositivo MICRO).	106
6.6. Resultados EER medio - Dominio de la Frecuencia (Dispositivo MOTO).	106
7.1. Visión Android Studio.	116
7.2. Opción USB conexión teléfono móvil-PC.	118
7.3. Depuración por Bluetooth Móvil-Reloj.	119
7.4. Menú aplicación WearableSensors desde el teléfono móvil.	120
7.5. Aplicación recogiendo datos	120

Índice de tablas

2.1. Gestión de riesgos	28
2.2. Plan de contingencia para la gestión de riesgos	30
2.3. Lista de tareas a realizar	35
3.1. Resultados de artículos utilizando un smartphone.	60
3.2. Resultados cambiando la posición del dispositivo.	62
4.1. Metadatos de los usuarios en la Base de Datos inicial.	64
6.1. Autocorrelaciones obtenidas para la muestra 1 y sesion 1 del usuario19	93
6.2. Resultados EER medio del Dominio del Tiempo (Dispositivo MICRO)	98
6.3. Resultados EER medio del Dominio del Tiempo (Dispositivo MOTO)	98
6.4. Resumen Mejores Resultados en Multisesión-Multimuestra - dominio del tiempo.	98
6.5. Tabla completa del módulo en el Dominio del Tiempo (Dispositivo MICRO)	101
6.6. Tabla completa del módulo en el Dominio del Tiempo (Dispositivo MOTO)	101
6.7. Combinación ACC/GYR del módulo en el Dominio del Tiempo.	102
6.11. Resumen Usuarios EER alto Combinación ACC/GYR - dominio del tiempo.	102
6.8. Combinación ACC/GYR de la componente X en el Dominio del Tiempo.	103
6.9. Combinación ACC/GYR de la componente Y en el Dominio del Tiempo.	103
6.10. Combinación ACC/GYR de la componente Z en el Dominio del Tiempo.	104
6.12. Resultados EER medio del Dominio de la Frecuencia (Dispositivo MICRO)	105
6.13. Resultados EER medio del Dominio de la Frecuencia (Dispositivo MOTO)	105
6.14. Resumen Mejores Resultados en Multisesión-Multimuestra - dominio de la frecuencia.	105
6.15. Tabla completa del módulo en el Dominio de la Frecuencia (Dispositivo MICRO)	108
6.16. Tabla completa del módulo en el Dominio de la Frecuencia (Dispositivo MOTO)	108
6.17. Combinación ACC/GYR del módulo en el Dominio de la Frecuencia.	109

6.18. Combinación ACC/GYR de la componente X en el Dominio de la Frecuencia. . .	110
6.19. Combinación ACC/GYR de la componente Y en el Dominio de la Frecuencia. . .	110
6.20. Combinación ACC/GYR de la componente Z en el Dominio de la Frecuencia. . . .	111
6.21. Resumen Usuarios EER alto Combinación ACC/GYR - Dominio de la Frecuencia.	111
6.22. Resumen de la combinación de sensores en el dispositivo MICRO.	112
6.23. Resumen de la combinación de sensores en el dispositivo MOTO.	112

Capítulo 1

Introducción

A medida que pasan los años, las tecnologías evolucionan, buscando hacer la vida de las personas más cómoda. Desde las comunicaciones, donde actualmente basta con acceder a un teléfono móvil u ordenador y pulsar unos botones, cuando antes se necesitaba comprar un sobre, un sello, escribir la carta y buscar un buzón de correos para enviarla, esperando que con suerte al día siguiente o a los dos días, el receptor recibiera la carta e hiciera el mismo procedimiento para poder conocer su respuesta. Pero también los procesos de compra, y no sólo de ropa, sino también de artículos frescos (fruta, verdura...) que aunque aún no está muy desarrollado, ya hay empresas que trabajan en mejorarlo, a través de la tecnología de visión artificial y las cámaras de alta tecnología de seguimiento vinculadas a cada sección del supermercado, para que el cliente online pueda coger su turno para ser atendido, informándole del tiempo de espera, y en el momento de iniciar su turno, le permitirá interactuar con el dependiente, ya sea para indicar el producto que desea y cómo quiere que se lo prepare como para solicitar recomendaciones o hacerle preguntas como si estuviera en la tienda física [1]. Pero también los cada vez más conocidos drones, que son vehículos aéreos no tripulados que se pueden manejar con control remoto o a través de aplicaciones para smartphones o tablets, ya se han utilizado para sobrevolar el mundial de fútbol en Brasil para mostrar desde arriba el juego con los mejores ángulos, como ninguna otra cámara podía o en Rusia y Brasil donde los drones se utilizan para hacer envíos de pizza, y aunque parezca extraño, los drones destacan por su efectividad en situaciones límite, pudiendo acceder a áreas que se quedan aisladas o de difícil acceso a una velocidad de vuelo que permite recorrer áreas enormes en muy poco tiempo, y pueden servir en situaciones de emergencia [2].

Cuando un usuario quiere llevar a cabo cualquier tipo de comunicación o transacción a través de Internet, necesita algún tipo de autenticación. Los sistemas más utilizados actualmente requieren que el usuario se registre y recuerde contraseñas, provocando que este termine realizando prácticas poco seguras, como emplear contraseñas simples, repetir una misma contraseña en varios sitios web, guardarlas en algún archivo, no cambiarlas con el paso del tiempo, etc. También hay que añadir problemas administrativos derivados de posibles pérdidas de claves, o en el caso de que el usuario necesite llevar consigo alguna tarjeta o dispositivo, esta también se puede perder, ser robada o transferirse.

Una alternativa es el uso de sistemas basados en biometría, puesto que el usuario no necesita recordar ni llevar consigo nada, empleando únicamente las características intrínsecas (físicas o de la forma de actuar) del usuario. La biometría se lleva aplicando desde finales del siglo XIX para la identificación de las personas con métricas como la huella dactilar, la cual sigue usándose hoy en día. Las mejoras tecnológicas y la necesidad de incrementar y simplificar la identificación de los usuarios ha provocado que a lo largo de los últimos años los estudios basados en sistemas de reconocimiento biométrico hayan cobrado una mayor relevancia.

Tradicionalmente, los métodos empleados en reconocimiento biométrico estaban relacionados con las características físicas del individuo como su cara o su huella dactilar. Estas características dan buenos resultados, pero requieren que el usuario ponga su cara o su dedo en algún dispositivo, proceso que termina siendo incómodo para el individuo. Por eso, en la actualidad y en este proyecto se busca emplear el comportamiento de la forma de actuar del individuo para el reconocimiento biométrico, por ser métodos menos intrusivos.

Se propone aprovechar el desarrollo y gran éxito en ventas en los últimos años de los dispositivos ponibles comerciales (pulseras de actividad, relojes inteligentes, etc.) para ver si es posible verificar a una persona a partir de sus datos. Estos dispositivos ya tienen aplicaciones muy diversas, desde recibir y contestar a notificaciones del teléfono móvil hasta monitorizar el rendimiento de los deportistas para ayudarles durante su entrenamiento. Gracias a que cuentan con una gran variedad de sensores, capaces de capturar distintos tipos de datos del usuario como su movimiento, ritmo cardíaco, sudor... hace posible su uso en el reconocimiento biométrico.

Teniendo en cuenta el tipo de sensores que incorporan actualmente los dispositivos ponibles comerciales, y como continuación de dos trabajos previos, se va a centrar la atención en verificar a una persona a partir de su *forma de andar*. En el primer trabajo previo [3] se hizo un estudio más detallado de todos los sensores de distintos dispositivos ponibles y se llegó a la conclusión de que los únicos que proporcionaban una información susceptible de ser usada en biometría para la forma de andar son los aquí se van a usar: el acelerómetro y el giroscopio. También se construyó una aplicación Android que permitía la recogida de los datos de dos dispositivos seleccionados. El segundo trabajo [4] recogió datos de diversos usuarios y realizó un estudio preliminar con ellos que demostraba la existencia de periodicidad en la señal de los datos y resultados positivos que indicaban su posible uso en biometría.

La diferencia entre la forma de abordar el problema en este proyecto y el resto de los trabajos de la bibliografía, es que se van a utilizar dispositivos comerciales, cuando en la bibliografía se utilizan dispositivos creados ad hoc para el propio propósito del proyecto o smartphones.

La biometría puede ser aplicada para la identificación o la autenticación de las personas. Los sistemas de verificación son configurados para cada usuario registrado con el objetivo de verificar la identidad de dicho usuario en una etapa posterior mientras que, en la identificación, el sistema presenta una señal biométrica y se debe decidir quién es el propietario de esa señal de entre un grupo de usuarios registrados. En otras palabras, en verificación se busca respuesta a la pregunta *Am I who I claim I am?* (*¿Soy yo quien digo que soy?*), mientras que la identificación busca

respuesta a la pregunta *Who am I?* (*¿Quién soy yo?*). En este proyecto, nos vamos a centrar únicamente en la verificación, cuyas fases se muestran, de manera resumida, en el esquema de la figura 1.1.



Figura 1.1: Fases involucradas en el sistema de verificación biométrica.

El sensor captura la señal biométrica cruda, se realiza una fase de preprocesamiento y extracción de características de las señales, después [...] indica la generación de los ficheros de salida necesarios para aplicar el clasificador, el cual generará una métrica, que servirá para evaluar el rendimiento del sistema construido, y posteriormente podrá ser utilizada para tomar una decisión: soy yo o no soy yo el usuario. Este trabajo se ha realizado de manera conjunta con otro del Grado en Estadística [5], donde se ha centrado el trabajo en la “Unidad de procesamiento”, haciendo un análisis profundo de la etapa de preprocesamiento y sus distintos parámetros, así como estudiando distintos tipos de conjuntos de características y haciendo una selección de las mismas. En el presente proyecto se van a utilizar esas decisiones para construir un sistema de reconocimiento final, el cual se va a evaluar para los sensores y dispositivos que se tienen disponibles, intentando poner en marcha y capturar nuevos datos de individuos, si el tiempo lo permite.

La Base de Datos a utilizar va a ser la obtenida en el Trabajo Fin de Grado (TFG) previo a este [4]. El hecho de que el tamaño de la base de datos no sea muy grande, no es ningún problema, ya que el objetivo de este trabajo no es obtener un sistema de reconocimiento basado en ponibles, si no profundizar en el conocimiento y las especiales características de esta biometría, trabajo no realizado hasta ahora en la bibliografía. Además, el hecho de ser una base de datos no muy numerosa hace que sea factible un análisis particular del comportamiento de cada individuo, lo que, como se verá, nos va a permitir extraer conclusiones muy interesantes para trabajos futuros.

Se ha realizado un apartado con los conceptos comunes en biometría y los sistemas biométricos que aparecerán de manera recurrente a lo largo del trabajo, las cuestiones éticas sujetas a este proyecto y una breve discusión sobre la privacidad de los datos, pero por no repetir información, si se quiere conocer dicho estudio realizado, se puede ver en el TFG complementario a este [5].

1.1. Biometría

En este apartado, se van a explicar una serie de conceptos comunes en biometría y los sistemas biométricos que aparecerán de manera recurrente a lo largo del trabajo [6].

La **biometría** es el estudio estadístico de los fenómenos o procesos biológicos. Tiene muchas aplicaciones posibles, pero dentro de las tecnologías de la información, la más destacada es el

estudio del reconocimiento de los seres humanos a partir de sus características, que se suelen clasificar en dos tipos:

- **Características fisiológicas:** Son características físicas de los individuos. Dentro de este grupo cabe destacar la huella dactilar, el iris, etc. Se caracterizan por ser estáticas, es decir, no cambian con el tiempo.
- **Características del comportamiento:** Son propiedades de la forma de actuar de los individuos. Dentro de este grupo se encuentra el modo con el que interactúan con los dispositivos, su voz, firma, forma de andar, etc. Se caracterizan por ser dinámicas, es decir, pueden cambiar con el paso del tiempo.

Un **sistema de reconocimiento biométrico** es una aplicación informática con la capacidad de identificar o verificar a una persona a partir de sus características, bien sean fisiológicas o de comportamiento.

Los sistemas de reconocimiento necesitan algún tipo de patrón para poder identificar o verificar a los individuos. Un **patrón** es un modelo creado mediante capturas o datos del usuario para representarle.

Evidentemente, no todas las características de un individuo pueden ser empleadas para el reconocimiento biométrico. Según [7–9], para que una característica biométrica pueda ser considerada como tal, ésta ha de cumplir las siguientes propiedades.

- **Universalidad:** todas las personas han de tener dicha característica biométrica.
- **Unicidad:** no ha de haber dos personas que sean idénticas atendiendo únicamente a esa característica.
- **Permanencia:** o biológicamente constante, es decir, la característica no tiene que variar con el tiempo.
- **Recolectable:** la característica ha de poder ser medible cuantitativamente.

Buscando conseguir un sistema de reconocimiento biométrico que tenga las siguientes características.

- **Rendimiento:** precisión que tiene el sistema biométrico empleado a la hora de identificar o verificar a un individuo.
- **Aceptabilidad:** el grado en que el público se muestra positivo a utilizar el sistema biométrico.
- **Invulnerabilidad:** el grado de facilidad del sistema a ser engañado mediante el uso de técnicas fraudulentas.

En la biometría basada en comportamiento como la forma de andar, en ocasiones, la propiedad de *permanencia* no se cumple, denominando a este tipo de biometrías como suaves o débiles (*Soft Biometrics*).

Por otro lado, la posibilidad de verificar o identificar a un individuo a través de su forma de andar está sujeto a cuestiones éticas, teniendo una serie de ventajas e inconvenientes.

■ VENTAJAS

- No requiere interacción durante el proceso de verificación o identificación, el usuario simplemente tiene que andar.
- Reconocimiento continuo, el propietario se mantiene automáticamente autorizado para el acceso al dispositivo.
- Proceso discreto, sin molestar al usuario. No requiere cooperación explícita del sujeto.
- Se puede capturar la información a distancia.
- Podría utilizarse como ventaja en el campo de la asistencia sanitaria, detectando los cambios de la forma de andar para ayudar a identificar los primeros indicadores de la aparición de la enfermedad de Parkinson y la esclerosis múltiple, así como otras enfermedades.
- Un impostor puede observar cómo camina un usuario, pero aun así tendrá dificultades para replicar su patrón de marcha, es decir, es difícil de robar o falsificar.

■ INCONVENIENTES

- Existen factores externos que influyen en la forma de andar de las personas: condiciones de la superficie, meteorológicas, la ropa o los zapatos que lleve el usuario, etc.
- Existen factores internos que influyen en la forma de andar de las personas: estado físico, mental, una enfermedad, etc.

■ CUESTIONES ÉTICAS

- Los conjuntos de datos contienen información muy sensible, por poderse utilizar para identificar de forma única a las personas y dependiendo del tipo de sensor usado, se podría incluir información que pudiera revelar las condiciones médicas de los usuarios.
- No requiere el consentimiento del individuo que se está observando, por lo que se podría estar extrayendo su información sin que el usuario lo sepa.

La privacidad de los datos es un problema cada vez más presente en nuestra sociedad. Hay que tener mucho cuidado porque aunque aparentemente sólo estemos observando su forma de andar, puede existir gente que de manera maliciosa aproveche esa información y consiga conocer la identidad física, fisiológica o psíquica de los usuarios. Como se ha podido ver en las ventajas e inconvenientes, esta información se puede utilizar en el campo de la *medicina* de manera positiva, en la prevención

de enfermedades y/o la posible actuación temprana de las mismas, o de manera negativa, revelando las condiciones médicas de los usuarios y utilizandolo para perjudicarlos. Para intentar evitar problemas, ya existe la primera guía de pautas éticas para hacer el uso de la Inteligencia Artificial (AI) más responsable, producidas por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial (AI HLEG) a nivel de la Unión Europea (EU) [10]. En ella se reconoce el enorme impacto positivo que la AI tiene a nivel mundial, tanto comercial como socialmente, siendo una tecnología tanto transformadora como disruptiva que ha ido evolucionando en los últimos años produciendo enormes cantidades de datos digitales, creando una importante innovación científica y de ingeniería. Aseguran que la AI continuará impactando a la sociedad y a los ciudadanos de una manera que aún no podemos imaginar. Por ello, consideran importante que se preste la debida atención a garantizar un entendimiento y un compromiso para construir una AI digna de confianza y han redactado las directrices para que esto sea así, asegurando el propósito ético. Y aunque, afirman que la AI puede provocar daños no intencionados, han desarrollado un marco para implementar la AI confiable, ofreciendo una orientación concreta para su logro, proponiendo métodos técnicos y no técnicos de ayuda para su realización e implementación. De todas formas, en este trabajo se utilizará la información anonimizada.

1.2. Motivación

La elección de este estudio para llevar a cabo mi Trabajo de Fin de Grado fue debido a mi interés por el mundo de la biometría. Incluso antes de elegir la carrera que ahora termino, ya me llamaba la atención por parecerme un campo innovador en el que se había trabajado mucho a lo largo de los años, pero sobre el que todavía quedaba mucho trabajo por hacer.

Cuando mis tutores me presentaron este tema me pareció original, novedoso y con ese toque de dificultad que me llamaba la atención.

Entre las asignaturas del Grado en Ingeniería Informática que más me habían gustado a lo largo de estos cinco años y con las que había conseguido mi afición por los datos, se encontraban:

- **Técnicas de Aprendizaje Automático:** donde conseguí un conocimiento más profundo de mayor cantidad de técnicas de Machine Learning, trabajando con diferentes metodologías experimentales y aplicando los conocimientos teóricos en la resolución de prácticas.
- **Minería de Datos:** donde asenté y afiancé mis intereses al reforzar los conocimientos de todas las etapas del proceso de minería de datos, desde el preprocesamiento de los datos hasta la evaluación de los resultados, extrayendo conocimiento con diversas técnicas de aprendizaje más profundo, así como el conocimiento de los métodos de ensamblado.

No obstante, a lo largo de la titulación nunca había trabajado con datos biométricos ni conocía el mundo de la biometría, pero me parecía que existían muchas posibilidades con las que trabajar y eso hizo que me terminara de decidir.

1.3. Objetivos

A continuación, se indica el objetivo general y se enumeran los objetivos específicos en que se divide el presente proyecto.

1.3.1. Objetivo general

El objetivo principal de este trabajo es estudiar el posible uso de dispositivos ponibles comerciales para el reconocimiento biométrico de personas mediante la forma de andar.

1.3.2. Objetivos específicos

Para poder cumplir el objetivo general, se han creado una serie de objetivos específicos que se desarrollarán de forma progresiva. Estos objetivos servirán para determinar si merece la pena continuar futuros estudios en este tema o, por el contrario, si es mejor abandonar esta línea de investigación. Los objetivos específicos que se han planteado llevar a cabo han sido los siguientes:

1. Realizar un análisis de los datos para ver qué características tienen y en qué condiciones se encuentran.
2. Evaluar las diferencias entre el acelerómetro y el giroscopio.
3. Contrastar los resultados con respecto al tipo de ponible usado.
4. Si todo lo anterior genera resultados prometedores y se tiene el tiempo suficiente, recopilar datos de al menos 25 nuevos voluntarios para repetir el estudio y contrastar los resultados obtenidos.

Como ya se ha dicho, este trabajo se realiza de manera conjunta con otro del Grado en Estadística [5] donde los objetivos son complementarios a estos. Aquí se evalúan y contrastan las diferencias utilizando los sensores y dispositivos disponibles. Para ello se necesita la construcción de un sistema de reconocimiento, que es lo que se realiza en el TFG de Estadística.

1.4. Estructura de la obra

Esta memoria se encuentra dividida en una serie de capítulos y secciones basándose en la estructura presentada en la guía docente para la asignatura TFG del Grado en Ingeniería Informática de la *Universidad de Valladolid*.

De esta manera siguiendo al actual capítulo introductorio donde se habla de biometría, la motivación por este trabajo y los objetivos, se encuentran los siguientes capítulos.

Capítulo 2. Plan de desarrollo y seguimiento: En este capítulo se hablará sobre el modelo de proceso seguido para realizar este trabajo, de la planificación inicial y la real llevada a cabo indicando las tareas realizadas en cada una de las semanas y el software utilizado.

Capítulo 3. Estado del Arte: En este capítulo se analizarán los trabajos previos existentes en el campo de la biometría considerando los diferentes sensores existentes de manera breve y centrándose en los que se van a utilizar en el presente proyecto.

Capítulo 4. Datos y análisis: se explicarán los datos con los que se van a trabajar y se realizará un análisis inicial que incluirá únicamente la limpieza de los datos con los problemas que vayan surgiendo.

Capítulo 5. Configuración experimental: En este capítulo se explicarán los parámetros estáticos que se vayan a fijar en el sistema de reconocimiento final. Las decisiones serán sobre qué características se van a extraer de los datos y en qué dominios, cómo se van a medir los resultados, qué procedimiento experimental se va a seguir y qué algoritmo de clasificación se va a utilizar, justificando cada una de las decisiones.

Capítulo 6. Experimentos: Resultados finales: Con todas las decisiones tomadas, conseguidas utilizando un dispositivo y sensor concreto, se va a probar el sistema de reconocimiento final construido a las diferentes posibilidades de sensor/dispositivo/procedimiento experimental.

Capítulo 7. Adquisición de nuevos datos: Se pondrán en marcha los dispositivos que permiten la adquisición de nuevos datos, explicando cómo se ha conseguido.

Capítulo 8. Conclusiones y trabajo futuro: Este es el capítulo final donde se expondrán las conclusiones obtenidas y las posibles alternativas a probar en un futuro.

Para finalizar se encuentra una sección donde se explican los acrónimos y abreviaturas utilizadas a lo largo de la memoria, un índice alfabético, los anexos del trabajo y la bibliografía.

Capítulo 2

Plan de desarrollo y seguimiento

2.1. Modelo de proceso

Los esquemas tradicionales de desarrollo han demostrado ser efectivos en numerosas situaciones. Sin embargo, están orientados a proyectos de gran tamaño respecto a tiempo y recursos, no resultando los más adecuados para proyectos pequeños, en los que el entorno del sistema es muy cambiante.

Las metodologías ágiles constituyen una solución a medida para proyectos pequeños, aportando una gran simplicidad y al mismo tiempo, permitiendo llegar a soluciones finales de calidad. Sus principios permiten construir el proyecto en torno a individuos motivados y con mayor confianza al realizar las tareas.

El presente proyecto, formado por una única persona y tres tutores, va a seguir una adaptación del SCRUM original. Otros proyectos individuales también lo han hecho como, por ejemplo, las aplicaciones de Ten Kettles [11], desarrolladas en solitario por Alex Andrews, desarrollador que consigue trabajar de manera más productiva, con un crecimiento constante y una mayor felicidad, adoptando lo que se puede llamar *Scrum Of One* [12].

Aunque existe una limitación importante: el tiempo, fijado en un total de 450 horas con fecha de inicio y fin conocidas a priori e ilustradas a través de un diagrama de Gantt que puede verse en la figura 2.2, esto será meramente orientativo, prevaleciendo la utilización de la adaptación a SCRUM y lo planeado en las diferentes iteraciones.

Las razones que han llevado a la utilización de metodología ágil, a la que llamaremos *Light Scrum*, están basadas en sus principios [13–15]:

- Desarrollo iterativo e incremental.
- Autoorganización con responsabilidades y compromisos claros en cada iteración.
- Diálogo, comunicación y aprendizaje continuo: alumno-tutores.

- Flexibilidad ante los cambios.
- Rapidez cuando las cosas no están funcionando como estaba previsto al inicio del proyecto.
- A intervalos regulares (iteraciones) se reflexiona sobre cómo ser más efectivo para, a continuación, ajustar y mejorar el comportamiento en consecuencia.

En este proyecto, se van a fijar los siguientes parámetros:

- Duración de cada iteración: 2 semanas.
- Número de iteraciones: un máximo de 10.

A través de reuniones semanales, como se muestra en el calendario de la figura 2.1, se va a comunicar la información alumno-tutores de manera regular, eficiente y efectiva, ayudando a continuar en todo momento por buen camino, con la opción de equivocarse y poder reajustar el trabajo rápidamente. La primera iteración comienza el día 23 de enero, planeando la finalización del trabajo el día 29 de mayo, tras 9 iteraciones, o el 12 de junio, tras 10 iteraciones, en caso de complicaciones.

La finalización de una iteración, lo que en Scrum recibiría el nombre de *sprint*, va a conllevar las siguientes tareas:

- Comprobar si se han cumplido los objetivos planeados para la iteración.
 - En caso de que no se hayan cumplido, analizar por qué, qué es lo que ha fallado, y cómo se puede solucionar, así como otros fallos y sus correcciones que sean relevantes para que la siguiente iteración sea satisfactoria.
 - En caso de que sí se hayan cumplido, analizar cómo se ha trabajado y si se ha necesitado más tiempo.
- Comprobar con los tutores, en la reunión semanal, la evolución del proyecto.
- Planificar la siguiente iteración, en base a las tareas no completadas en la iteración anterior y las siguientes de la lista de la tabla 2.3.

2.2. Restricciones

Las restricciones asociadas al proyecto se deben a los siguientes factores que limitan las posibilidades de la solución final:

- Tiempo: limitado, 450 horas para la realización del proyecto completo.
- Equipo de trabajo: reducido, formado por una única persona y 3 tutores.

- Recursos: limitados, respecto a potencia de la computadora y la máquina virtual, así como la disponibilidad de datos y las posibilidades de recolección de nuevos datos.
- Conocimientos: limitados. A lo largo de la carrera hay muchas asignaturas y muy variadas, pero tanto el conocimiento de Android para el manejo de la aplicación para recoger datos, como la metodología utilizada en biometría y las técnicas de aprendizaje automático conocidas son limitadas y posiblemente insuficientes para afrontar un proyecto real de manera individual.

2.3. Gestión de riesgos

Según la guía de los fundamentos para la dirección de proyectos (del inglés A Guide to the Project Management Body of Knowledge o PMBOK [16]) la gestión de riesgos es una de las áreas más importantes que debe haber en la estructura de un proyecto.

A continuación, en la tabla 2.1, se enumera una lista de posibles riesgos, junto con su probabilidad de ocurrencia y el nivel de impacto que causaría sobre el proyecto, medidas ambas como una variable cualitativa ordinal que toma los niveles: alto, moderado, bajo o ninguno; y una estimación aproximada de la magnitud de la pérdida máxima en días.

ID	Riesgo	Probabilidad	Impacto	Magnitud de la pérdida (días)
1	Base de datos reutilizada del anterior TFG se encuentre vacía o incompleta.	Moderada	Alto	7
2	Perdida de los datos de la base de datos.	Moderada	Alto	2
3	Recursos insuficientes disponibles para realizar el trabajo (potencia de ordenador para procesar los datos).	Moderada	Alto	5
4	Caída de servidores de la Universidad de Valladolid (UVa): campus virtual, máquinas virtuales, y otros similares	Alta	Bajo	2
5	Pérdida de la conexión a internet.	Alta	Moderado	1
6	Caída de los servidores de Overleaf.	Alta	Moderado	1
7	Enfermedad o accidente que invalide a la autora del proyecto.	Moderada	Alto	5
8	Enfermedad o accidente que invalide a alguno de los tutores del proyecto.	Moderada	Alto	5
9	Estropearse el ordenador.	Alta	Moderado	3
10	Corte de luz.	Moderada	Moderado	1

ID	Riesgo	Probabilidad	Impacto	Magnitud de la pérdida (días)
11	Lentitud en la toma de decisiones, o decisiones incorrectas.	Alta	Alto	8
12	Retrasos excesivos en las iteraciones (sprints en SCRUM).	Alta	Alto	5
13	Incorrecta aplicación de la metodología ágil (Light Scrum).	Moderada	Moderado	3
14	Baja motivación.	Moderada	Alto	3
15	Requisitos mal entendidos o incompletos.	Alta	Alto	5
16	Complejidad elevada de las tareas.	Alta	Alto	5
17	Modificación de los requisitos.	Alta	Alto	5
18	Desarrollo incorrecto de la funcionalidad.	Alta	Alto	5
19	Insuficientes datos en la base de datos reutilizada para poder extraer buenas conclusiones.	Moderada	Alto	15
20	No funcionamiento de la aplicación desarrollada en TFGs anteriores para la recogida de datos.	Alta	Alto	15
21	Imposibilidad de quedar con los tutores, por incompatibilidad de horarios.	Baja	Alto	5

Tabla 2.1: Gestión de riesgos

Para cada uno de los riesgos enumerados en la tabla 2.1, se desglosa una posible manera de actuar que se adoptará en caso de ocurrencia. Los identificadores en la tabla de los riesgos y en la tabla 2.2 de planes de contingencia son los mismos.

ID	Plan de contingencia
1	Ponerse en contacto con el autor de la base de datos y los profesores para intentar solventar el problema en la mínima cantidad de tiempo posible. Si no hubiese solución, adoptar el plan de contingencia del riesgo 20.
2	Adoptar el plan de contingencia del riesgo 1.
3	Utilizar ordenadores proporcionados por la Escuela de Ingeniería Informática (Universidad de Valladolid). En caso necesario, ponerse en contacto con los tutores y los técnicos de la Escuela para asignar una máquina virtual.
4	En caso de caída del Campus Virtual, se contactaría con los profesores responsables a través del correo electrónico; en cuanto a la máquina virtual (si se utilizará), se realizarán copias de seguridad semanales de los datos de la base de datos.

ID	Plan de contingencia
5	En caso de no tener internet se desplazaría a otra ubicación donde sí hubiera, como bibliotecas o cibercafés, para realizar las tareas pertinentes.
6	Se realizará una copia de seguridad en local después de cada modificación de la memoria o cualquier archivo relacionado que se aloje en Overleaf.
7	Se dejará tiempo suficiente para la recuperación de la autora, retomando el trabajo cuando sea posible.
8	Se seguirá trabajando en el proyecto, resolviendo los problemas con el otro tutor. Dejando tiempo suficiente para la recuperación del tutor correspondiente y retomando el trabajo al completo cuando sea posible.
9	De forma similar a la pérdida de conexión a internet, el afectado deberá buscar otra computadora para realizar el trabajo.
10	Desplazamiento a un lugar donde haya corriente eléctrica o esperar a que ésta vuelva.
11	Analizar la causa de este retraso o fallo para evitarlo en ocasiones futuras. Podría plantearse tomar decisiones más “arriesgadas” e invertir menos tiempo en ellas, para ver si de esa forma se consiguen los mismos resultados.
12	La solución pasa por aumentar las horas de trabajo.
13	La Scrum Master deberá formarse mejor para intentar corregir la incorrecta aplicación de la metodología ágil (Light Scrum) y poder aprovechar al máximo sus beneficios.
14	Lidiar con la situación y buscar una solución para intentar aumentar los ánimos. Podrían realizarse sesiones de trabajo más dinámicas con los tutores.
15	Analizar el problema (requisito) con más cuidado, dedicándole más tiempo y preguntando las dudas a los profesores. Incrementar la carga de trabajo para adecuarla a la necesaria para cumplir con la planificación.
16	Realizar la tarea en el tiempo que se necesite e incrementar la carga de trabajo en las semanas siguientes para recuperar el tiempo perdido.
17	Dedicar tiempo en comprender los cambios y analizarlos en detalle para valorar las modificaciones, incrementando la carga de trabajo y organizando reuniones con los clientes si fuera necesario.
18	Dedicar más tiempo al desarrollo de dicha funcionalidad y pedir ayuda a los profesores, para conseguir terminar con éxito en la mínima cantidad de tiempo.
19	Puesta en marcha de la herramienta de adquisición de datos realizada en TFGs anteriores para conseguir nuevos y/o más datos.
20	Ponerse en contacto con los autores de la aplicación. En caso de no ser suficiente, hablar con profesores con los conocimientos necesarios para poner en marcha la aplicación (conocimiento de Android).
21	Intentar buscar otros horarios o hablar a través de correos electrónicos y/o Skype. Dedicar el tiempo de espera de la respuesta en organizar las cosas que se están haciendo y aquellas que queden por hacer, así como a escribir la memoria.

ID	Plan de contingencia
-----------	-----------------------------

Tabla 2.2: Plan de contingencia

2.4. Planificación

El TFG del programa de estudios conjunto de Grado en Ingeniería Informática y Grado en Estadística (INdat) está compuesto por el TFG de Ingeniería Informática de 12 European Credit Transfer and Accumulation System (ECTS) y el TFG de Estadística de 6 ECTS, lo que conlleva un total de 450 horas de dedicación. Por ser, en este caso, un trabajo conjunto de 18 ECTS. Aunque en este trabajo sólo se muestra la parte de informática, la planificación se ha hecho de manera común que es, en definitiva, como se ha trabajado.

El planteamiento básico inicial del trabajo se desglosa de la siguiente manera:

Haciendo referencia a días como cualquier día de la semana: de lunes a sábado. El sábado es opcional, para cuando se tengan muchos trabajos de las otras asignaturas, de manera general se intentarán realizar dichas horas entre el lunes y el viernes.

1. Una fase inicial de lectura, análisis y comprensión de documentos relacionados con el tema que se pretende abordar. Así como la adquisición y comprensión de la base de datos con la que se va a trabajar, que se calcula que llevará 2 días.
2. Búsqueda de artículos y documentación de la bibliografía más relevante, así como su lectura. Inicialmente llevará 15 días, pero se mantendrá presente durante todo el proyecto, por la necesidad de investigar sobre la tecnología actual más utilizada en las diferentes fases: preprocesamiento de los datos, extracción y selección de características, aplicación de clasificadores con diferentes metodologías experimentales e implementación de sistemas de reconocimiento.
3. Análisis visual de los datos que llevará unos 3 días, y que se hará al mismo tiempo que la fase anterior.
4. Estudio preliminar de los datos que agrupará:
 - Aplicación de técnicas de preprocesamiento: normalización, interpolación, filtros *low-pass/highpass*...
 - Visualización de los datos en el dominio del tiempo (DT) y de la frecuencia (DF).
 - Extracción de las zonas de interés.
 - Análisis de autocorrelación de una señal consigo misma.

El conjunto de todo ello llevará 35 días.

5. Extracción de características de los datos en el dominio del tiempo y de la frecuencia, que llevará 4 días.
6. Selección de las características de interés, que llevará 5 días y se realizará tras terminar la fase anterior.
7. Construcción de las tablas de los datos que contendrán toda la información con la estructura final deseada que se introducirá al clasificador. Esto englobará 5 días y se realizará de manera conjunta con las fases de extracción y selección de características.
8. Estudio prospectivo de los datos con implementación de algunas técnicas de clasificación y visualización de los resultados con las métricas que se consideren apropiadas, englobando un total de 4 días.
9. Desarrollo del diseño experimental del proyecto, que llevará 2 días.
10. Implementación de distintas alternativas de clasificadores, que llevará 4 días.
11. Análisis de los resultados, que llevará 8 días, pudiendo modificar algunas de las tareas anteriores para la mejora del sistema final.
12. Una fase final para completar la memoria de 15 días.
13. Realización y preparación de la presentación oral para la defensa final de 5 días.

En total resultan 107 días, comprendidos entre el 23 de enero y el 27 de mayo. Como 450 horas en 107 días son 4,21 horas/día, se va a redondear a 4 horas/día, posponiendo la finalización al día 3 de junio. Dejando bastantes días de margen antes de la fecha de cierre de solicitud de defensa en la convocatoria ordinaria (27 de junio) para posibles problemas.

Se dispone de dos dispositivos ponibles: pulsera *Microsoft Band 2* (MICRO) y reloj *Motorola Moto 360* (MOTO). En cada uno de ellos se han capturado datos del acelerómetro (ACC) y giroscopio (GYR), por lo que las tareas se realizarán en el siguiente orden:

1. Pulsera Microsoft y acelerómetro.
2. Pulsera Microsoft y giroscopio.
3. Reloj Motorola y acelerómetro.
4. Reloj Motorola y giroscopio.

Uniendo y comparando, en todos los casos, cada resultado con los anteriores.

A continuación, en la tabla 2.3, se listan las diferentes tareas, por orden de prioridad, estimando el esfuerzo en horas/hombre. La diferencia con las anteriores tareas reside en que en la tabla 2.3

se encuentran sobrevaloradas, habiendo más cosas en los últimos elementos que se realizarán únicamente si el tiempo lo permite.

El identificador ID es único para cada tarea a realizar. No obstante, las tareas se repiten entre los diferentes dispositivos, por lo que se ha indicado un número de tarea (Nº tarea) que muestra la cantidad de actividades diferentes que se van a realizar.

ID	Nº tarea	Lista de tareas del proyecto completo	Esfuerzo (h)
1	1	Lectura, análisis y comprensión de la memoria TFG del proyecto previo a este trabajo, el cual es continuación.	3
2	2	Adquisición y comprensión de la base de datos de los trabajos anteriores con la que se trabajará.	2
3	3	Búsqueda de artículos y documentación de la bibliografía más relevante.	10
4	4	Lectura y comprensión de artículos y documentación sobre trabajos similares.	35
5	5	Análisis y puesta en marcha de la herramienta de visualización realizada y utilizada en el TFG anterior.	1
Microsoft ACC			
6	6	Análisis visual de los datos.	6
7	7	Preprocesamiento de los datos: normalización, interpolación, filtros lowpass/highpass	6
8	8	Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).	7
9	9	Análisis de estabilidad, tendencia, estacionalidad, etcétera de los datos.	1
10	10	Análisis de autocorrelaciones de los datos.	3
11	11	Extracción de las zonas de interés de los datos (de manera manual).	45
12	12	Implementación de distintas alternativas para la extracción de características a partir de los datos (dominio del tiempo y de la frecuencia).	6
13	13	Análisis estadístico de las distintas características extraídas para estudiar cuáles son más discriminantes (selección de las características de interés)	20
14	14	Construcción de las tablas de los datos con la información estructurada de entrada al clasificador	6
15	15	Estudio prospectivo de los datos implementando algún clasificador y visualizando los resultados con las métricas necesarias.	20
16	16	Desarrollo del diseño experimental del proyecto	3

ID	Nº tarea	Lista de tareas del proyecto completo	Esfuerzo (h)
Microsoft GYR			
17	6	Análisis visual de los datos.	2
18	7	Preprocesamiento de los datos: normalización, interpolación, filtros lowpass/highpass	1
19	8	Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).	1
20	9	Análisis de estabilidad, tendencia, estacionalidad, etcétera de los datos.	1
21	10	Análisis de autocorrelaciones de los datos.	1
22	11	Extracción de las zonas de interés de los datos (de manera manual).	25
23	12	Implementación de distintas alternativas para la extracción de características a partir de los datos (dominio del tiempo y de la frecuencia).	1
24	13	Análisis estadístico de las distintas características extraídas para estudiar cuáles son más discriminantes (selección de las características de interés)	1
25	14	Construcción de las tablas de los datos con la información estructurada de entrada al clasificador	0,5
26	15	Estudio prospectivo de los datos implementando algún clasificador y visualizando los resultados con las métricas necesarias.	1
27	3	Búsqueda de artículos y documentación de la bibliografía más relevante.	3
28	4	Lectura y comprensión de artículos y documentación sobre trabajos similares.	4
Motorola ACC			
29	6	Análisis visual de los datos.	1
30	7	Preprocesamiento de los datos: normalización, interpolación, filtros lowpass/highpass	0,5
31	8	Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).	1
32	9	Análisis de estabilidad, tendencia, estacionalidad, etcétera de los datos.	0,5
33	10	Análisis de autocorrelaciones de los datos.	0,5
34	11	Extracción de las zonas de interés de los datos (de manera manual).	25

ID	Nº tarea	Lista de tareas del proyecto completo	Esfuerzo (h)
35	12	Implementación de distintas alternativas para la extracción de características a partir de los datos (dominio del tiempo y de la frecuencia).	2
36	13	Análisis estadístico de las distintas características extraídas para estudiar cuáles son más discriminantes (selección de las características de interés)	2
37	14	Construcción de las tablas de los datos con la información estructurada de entrada al clasificador	1
38	15	Estudio prospectivo de los datos implementando algún clasificador y visualizando los resultados con las métricas necesarias.	3
Motorola GYR			
39	6	Análisis visual de los datos.	1
40	7	Preprocesamiento de los datos: normalización, interpolación, filtros lowpass/highpass	0,5
41	8	Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).	1
42	9	Análisis de estabilidad, tendencia, estacionalidad, etcétera de los datos.	0,5
43	10	Análisis de autocorrelaciones de los datos.	0,5
44	11	Extracción de las zonas de interés de los datos (de manera manual).	25
45	12	Implementación de distintas alternativas para la extracción de características a partir de los datos (dominio del tiempo y de la frecuencia).	1
46	13	Análisis estadístico de las distintas características extraídas para estudiar cuáles son más discriminantes (selección de las características de interés)	2
47	14	Construcción de las tablas de los datos con la información estructurada de entrada al clasificador	0,5
48	15	Estudio prospectivo de los datos implementando algún clasificador y visualizando los resultados con las métricas necesarias.	1
49	16	Desarrollo del diseño experimental del proyecto	8
50	17	Pruebas de distintas alternativas de clasificadores.	2
51	18	Análisis de los resultados, buscando el mejor sistema.	25
52	3	Búsqueda de artículos y documentación de la bibliografía más relevante.	2

ID	Nº tarea	Lista de tareas del proyecto completo	Esfuerzo (h)
53	4	Lectura y comprensión de artículos y documentación sobre trabajos similares.	4
54	19	Extracción de las zonas de interés de los datos (de manera automática).	5
55	20	Implementación de un sistema de reconocimiento de manera automática.	35
56	21	Realización de la memoria. Incluye versiones y revisiones.	45
57	22	Puesta en marcha de la herramienta de adquisición realizada en TFGs anteriores.	3
58	23	Adquisición de más datos.	15
59	24	Estudio de lo realizado para los nuevos datos.	4
60	21	Realización de la memoria. Incluye versiones y revisiones.	5
61	25	Realización de la presentación para la defensa oral.	12
TOTAL			450

Tabla 2.3: Lista de tareas a realizar

Cabe la posibilidad de sustituir las tareas 22 y 23 (identificador 57 y 58) por la búsqueda de una base de datos pública de cualquier otro proyecto de investigación. Siempre y cuando los datos sean comparables con los usados en este trabajo y la información éste obtenida de manera similar, ya que sería mucho más rápido y se podría probar lo realizado (tarea 24) y obtener conclusiones más realistas.

En la figura 2.1 [17] se muestra el calendario con la fecha de inicio y fin previstas (color naranja), la fecha de cierre de solicitudes de defensa de TFGs (color rojo), los días de descanso (color gris oscuro), el periodo de vacaciones de Semana Santa, en el cuál sí se trabajará pero no se harán reuniones presenciales, y las reuniones semanales con los tutores del Grado en Ingeniería Informática (color amarillo). En la parte inferior se pueden ver los días totales que, aproximadamente, se van a dedicar cada mes.

Por otro lado, en la figura 2.2 se muestra el diagrama de Gantt completo con toda la planificación desde el principio hasta el final (23 de enero-3 de junio). Se han creado las tareas globales: Microsoft ACC, Microsoft GYR, Motorola ACC y Motorola GYR correspondientes a los diferentes dispositivos y cuyas subtareas corresponden a las mismas de la tabla 2.3. Se muestran en las figuras 2.3, 2.4, 2.5 y 2.6 donde se puede ver como tareas similares en los diferentes dispositivos van requiriendo cada vez menos tiempo.

Se dispone de un único recurso llamado I, correspondiente a la autora del presente TFG con una capacidad del 100% y sin sobreasignación en ningún momento, como se muestra en la figura 2.7. En esa figura se puede ver la dedicación total y mensual, junto con el tiempo planeado para invertir

2019

	enero	febrero	marzo	abril	mayo	junio	julio	
jueves								jueves
viernes								viernes
sábado								sábado
domingo								domingo
lunes							01	lunes
martes	01					01	02	martes
miércoles	02	REUNIÓN	REUNIÓN	REUNIÓN	REUNIÓN	REUNIÓN	03	miércoles
jueves	03						04	jueves
viernes	04						05	viernes
sábado	05						06	sábado
domingo	06						07	domingo
lunes	07						08	lunes
martes	08						09	martes
miércoles	09	REUNIÓN	REUNIÓN	REUNIÓN	REUNIÓN		10	miércoles
jueves	10						11	jueves
viernes	11						12	viernes
sábado	12						13	sábado
domingo	13						14	domingo
lunes	14						15	lunes
martes	15						16	martes
miércoles	16						17	miércoles
jueves	17						18	jueves
viernes	18						19	viernes
sábado	19						20	sábado
domingo	20						21	domingo
lunes	21						22	lunes
martes	22						23	martes
miércoles	23						24	miércoles
jueves	24						25	jueves
viernes	25						26	viernes
sábado	26						27	sábado
domingo	27						28	domingo
lunes	28						29	lunes
martes	29						30	martes
miércoles	30						31	miércoles
jueves								jueves
viernes								viernes
	8 días	24 días	26 días	26 días	27 días	2 días		

Figura 2.1: Calendario de desarrollo del proyecto propuesto inicialmente.

en cada tarea individual. Cuando aparece un recurso de I[50%] implica que se están realizando 2 tareas al mismo tiempo, ya que en ocasiones se dividen las tareas, sumando todas ellas la duración total planeada.

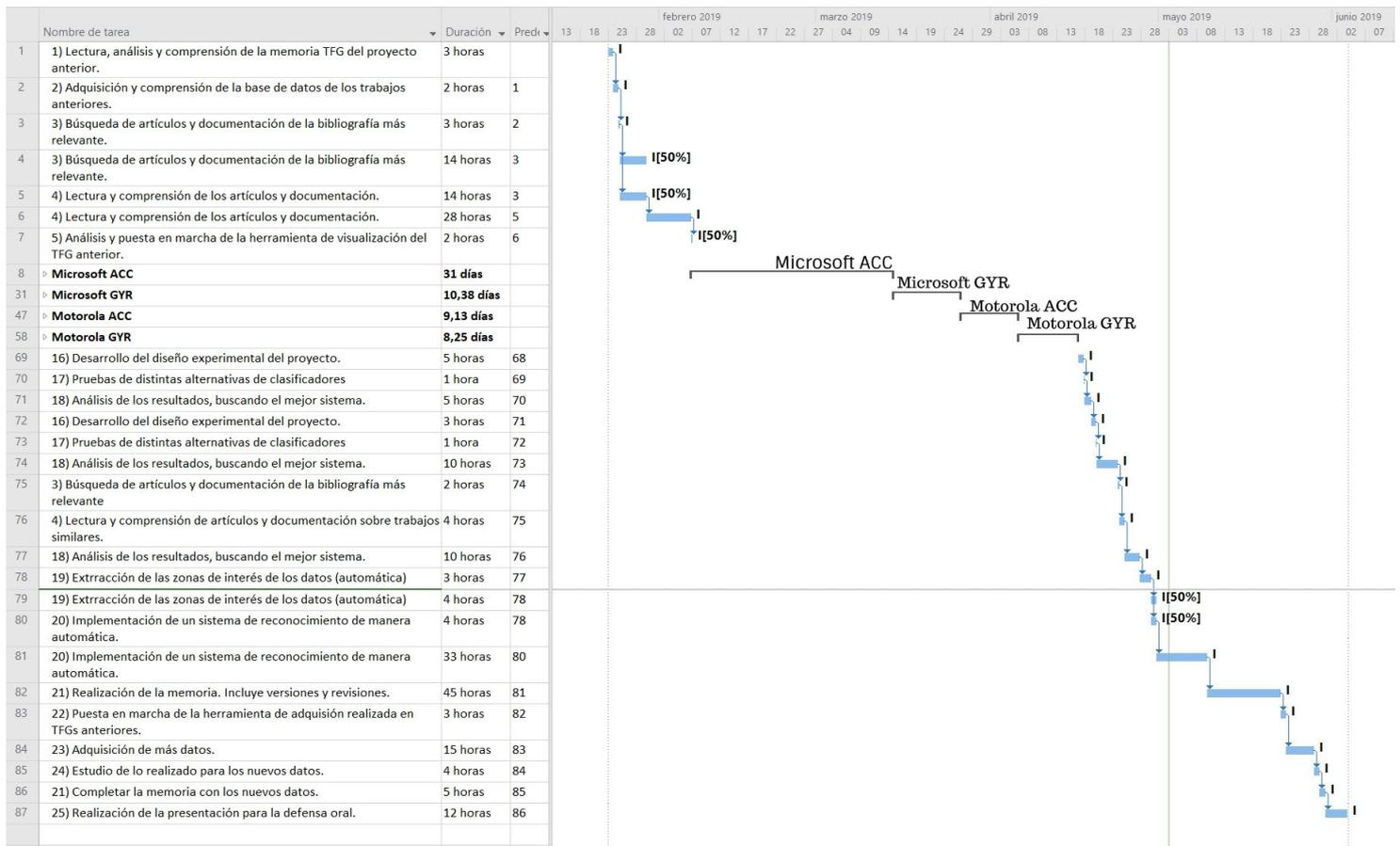


Figura 2.2: Diagrama de Gantt con la planificación completa.

2.5. Planificación de las iteraciones y seguimiento

En esta sección se mostrará la planificación de cada una de las iteraciones junto con su breve seguimiento a la finalización de cada una de ellas, tal y como se explica en la sección 2.1.

2.5.1. Iteración 1

Un TFG es un proyecto grande, completamente nuevo y diferente a los demás trabajos realizados hasta el momento en el resto de las asignaturas de la Universidad, por ser individual, de un tema novedoso y una duración relativamente larga. Por estas razones, esta primera iteración se dedicó a las tareas con identificador (ID) 1, 2, 3, 4 y 5.

1. Lectura, análisis y comprensión de la memoria TFG del proyecto previo a este trabajo, el cual es continuación.

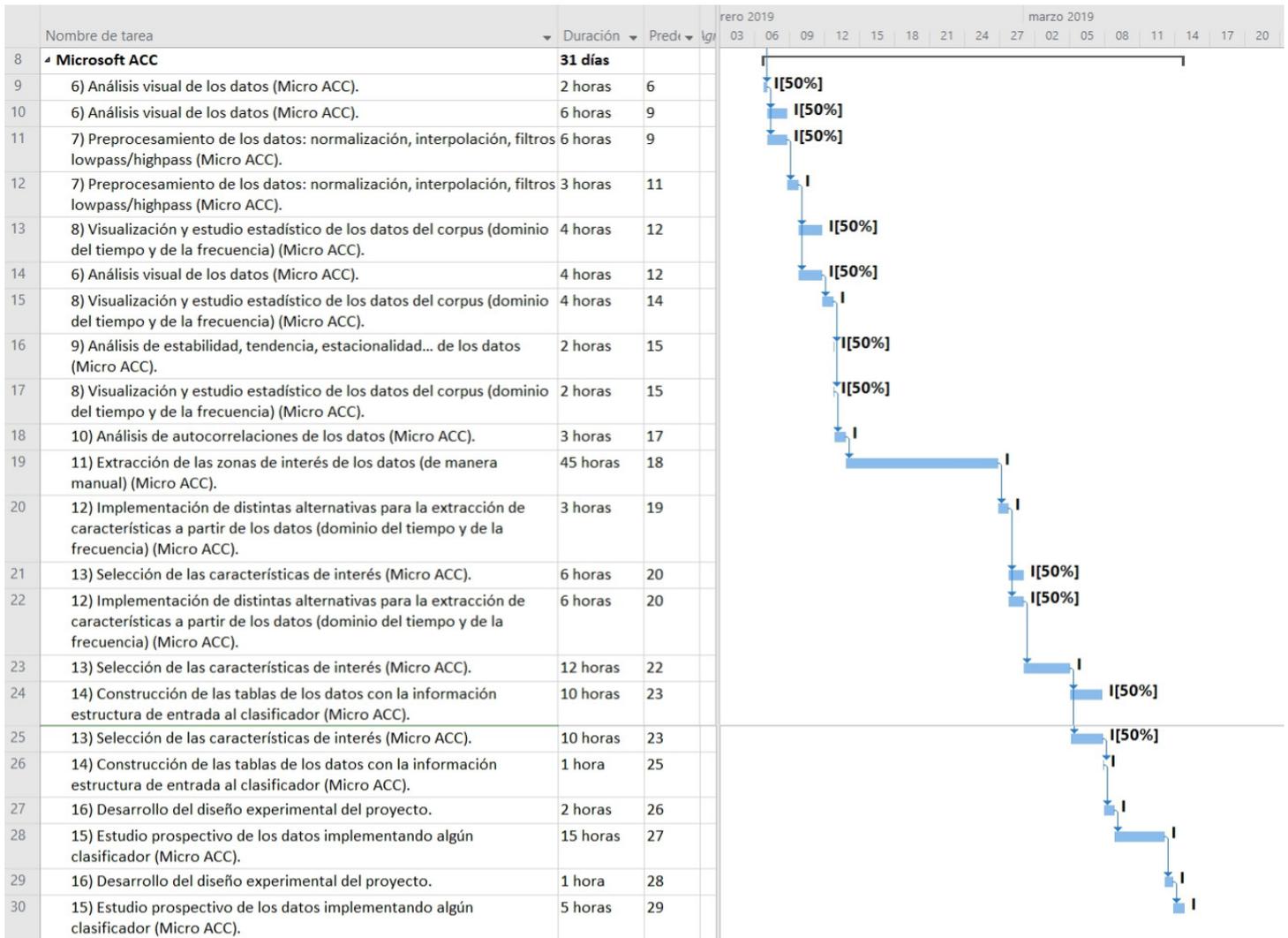


Figura 2.3: Subdiagrama de Gantt con la planificación del dispositivo Microsoft acelerómetro.

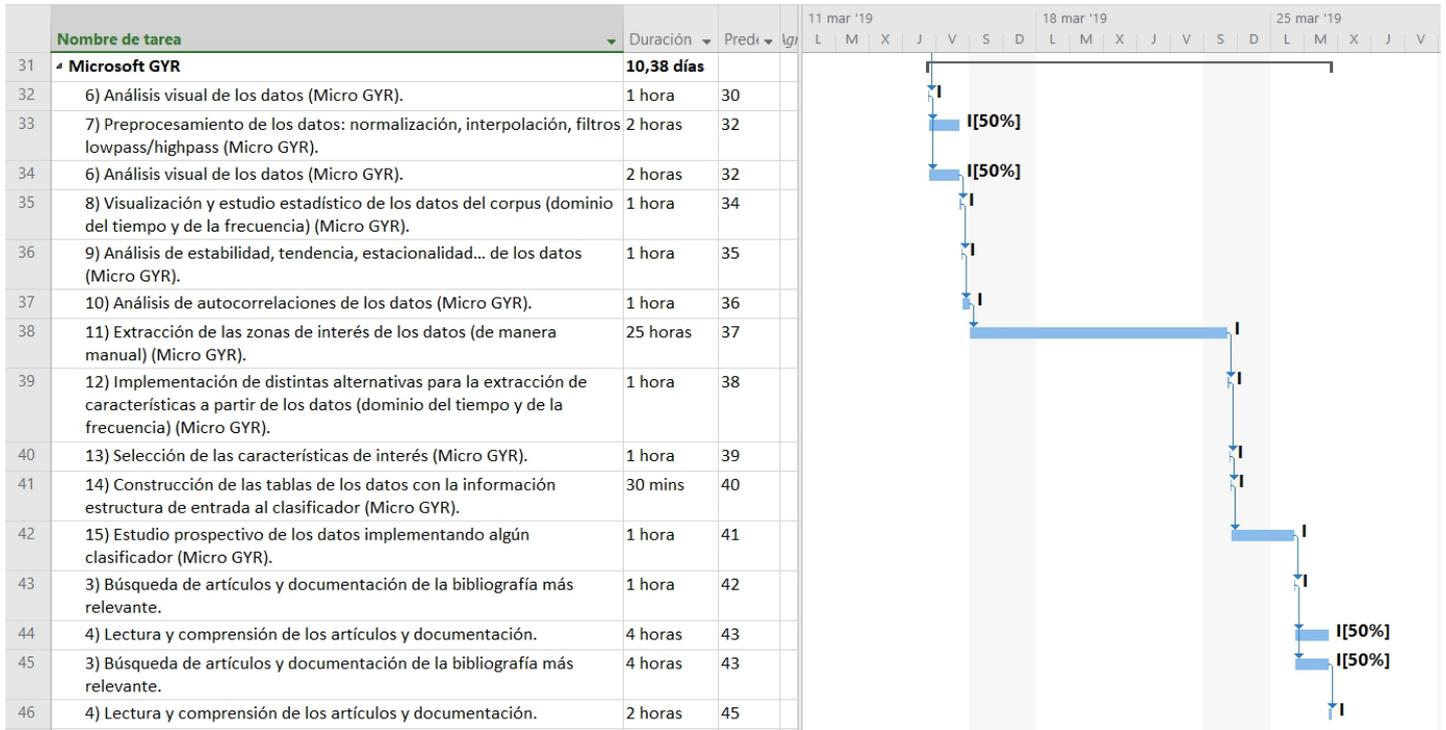


Figura 2.4: Subdiagrama de Gantt con la planificación del dispositivo Microsoft giróscopo.

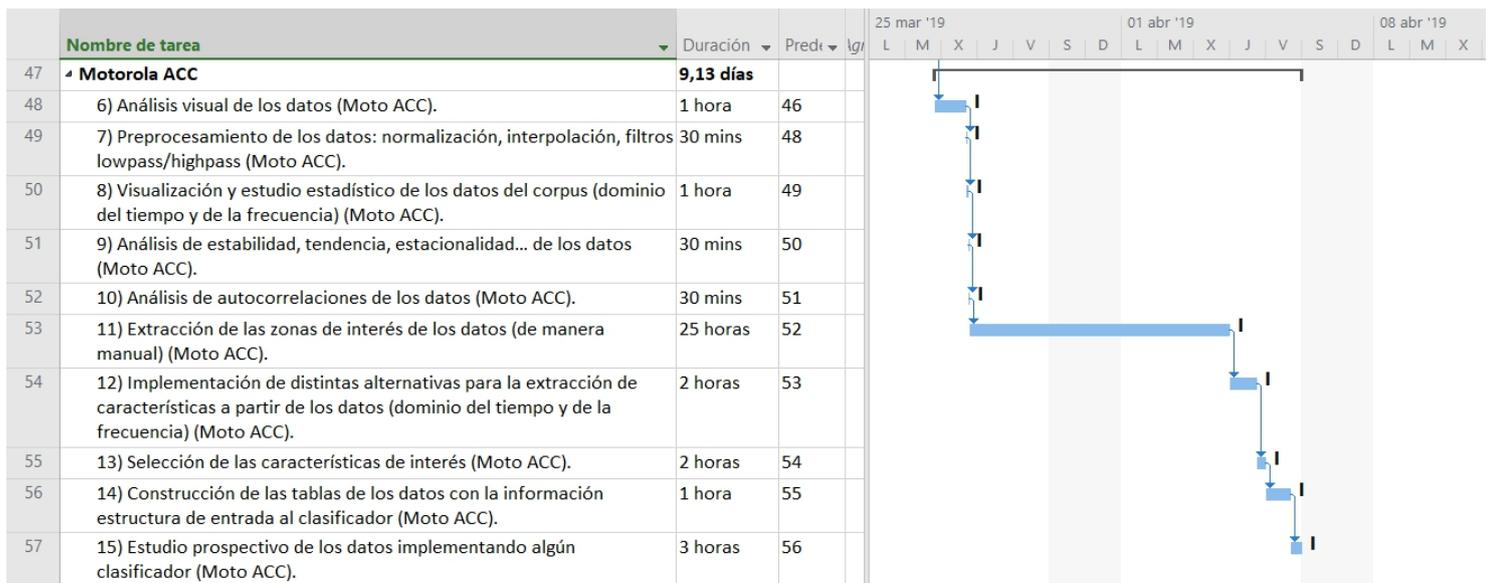


Figura 2.5: Subdiagrama de Gantt con la planificación del dispositivo Motorola acelerómetro.

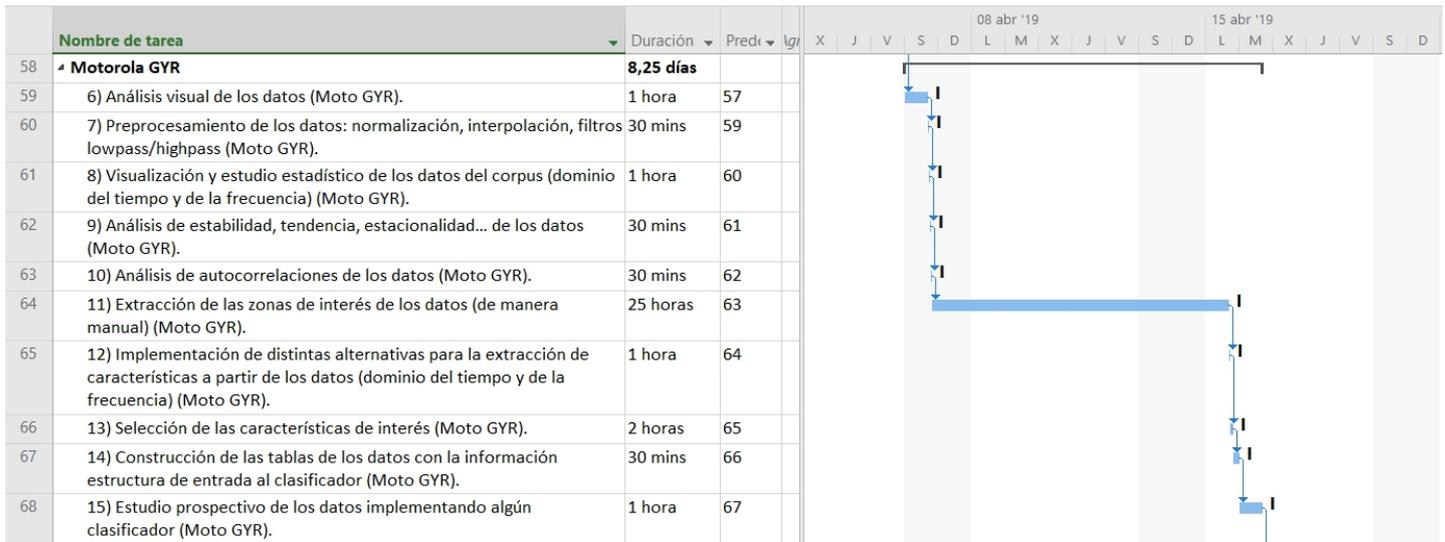


Figura 2.6: Subdiagrama de Gantt con la planificación del dispositivo Motorola giróscopo.

2. Adquisición y comprensión de la base de datos de los trabajos anteriores con la que se trabajará.
3. Búsqueda de artículos y documentación de la bibliografía más relevante.
4. Lectura y comprensión de artículos y documentación sobre trabajos similares.
5. Análisis y puesta en marcha de la herramienta de visualización realizada y utilizada en el TFG anterior [18].

La fecha de comienzo fue el miércoles 23 de enero, lo que llevó a su finalización el miércoles 6 de febrero. Tras estas dos semanas se cumplieron los ítems 1, 2 y 5 completamente y 3 y 4 a medias. No obstante, se realizó la búsqueda de los artículos (ítem 3) pero fue pequeña, ya que no estaba acostumbrada y la descarga tenía que ser desde la Universidad con conexión a Eduroam, que dificultaba el tiempo de búsqueda. También se leyeron algunos artículos (ítem 4) pero se dejó su finalización para la siguiente iteración. Se realizó el ítem 5 por su facilidad, ya que se disponía de la página de GitHub de Daniel González Alonso [18] donde se encuentra la herramienta de visualización junto con un breve vídeo explicativo que facilitaba su aprendizaje y uso.

2.5.2. Iteración 2

Para esta segunda iteración se planificaron los ítems:

3. Búsqueda de artículos y documentación de la bibliografía más relevante.
4. Lectura y comprensión de artículos y documentación sobre trabajos similares.
6. Análisis visual de los datos del acelerómetro en la pulsera de Microsoft.

Nombre del recurso	Trabajo	Detalles	1er trimestre			2º trimestre			may	jun
			dic	ene	feb	mar	abr			
4 I	450 horas	Trabajo			32h	96h	104h	104h	108h	6h
1) Lectura, análisis y comprensión	3 horas	Trabajo			3h					
2) Adquisición y comprensión de	2 horas	Trabajo			2h					
3) Búsqueda de artículos y doc	3 horas	Trabajo			3h					
3) Búsqueda de artículos y doc	7 horas	Trabajo			7h					
4) Lectura y comprensión de lc	7 horas	Trabajo			7h					
4) Lectura y comprensión de lc	28 horas	Trabajo			10h	18h				
5) Análisis y puesta en marcha	1 hora	Trabajo				1h				
6) Análisis visual de los datos (1 hora	Trabajo				1h				
6) Análisis visual de los datos (3 horas	Trabajo				3h				
7) Preprocesamiento de los da	3 horas	Trabajo				3h				
7) Preprocesamiento de los da	3 horas	Trabajo				3h				
8) Visualización y estudio esta	2 horas	Trabajo				2h				
6) Análisis visual de los datos (2 horas	Trabajo				2h				
8) Visualización y estudio esta	4 horas	Trabajo				4h				
9) Análisis de estabilidad, tena	1 hora	Trabajo				1h				
8) Visualización y estudio esta	1 hora	Trabajo				1h				
10) Análisis de autocorrelaci	3 horas	Trabajo				3h				
11) Extracción de las zonas de	45 horas	Trabajo				45h				
12) Implementación de distint	3 horas	Trabajo				3h				
13) Selección de las caracterís	3 horas	Trabajo				3h				
12) Implementación de distint	3 horas	Trabajo				3h				
13) Selección de las caracterís	12 horas	Trabajo					12h			
14) Construcción de las tablas	5 horas	Trabajo					5h			
13) Selección de las caracterís	5 horas	Trabajo					5h			
14) Construcción de las tablas	1 hora	Trabajo					1h			
16) Desarrollo del diseño expe	2 horas	Trabajo					2h			
15) Estudio prospectivo de los	15 horas	Trabajo					15h			
16) Desarrollo del diseño expe	1 hora	Trabajo					1h			
15) Estudio prospectivo de los	5 horas	Trabajo					5h			
6) Análisis visual de los datos (1 hora	Trabajo					1h			
7) Preprocesamiento de los da	1 hora	Trabajo					1h			
6) Análisis visual de los datos (1 hora	Trabajo					1h			
8) Visualización y estudio esta	1 hora	Trabajo					1h			
9) Análisis de estabilidad, tena	1 hora	Trabajo					1h			
10) Análisis de autocorrelaci	1 hora	Trabajo					1h			
11) Extracción de las zonas de	25 horas	Trabajo					25h			
12) Implementación de distint	1 hora	Trabajo					1h			
13) Selección de las caracterís	1 hora	Trabajo					1h			
14) Construcción de las tablas	0,5 horas	Trabajo					0,5h			
15) Estudio prospectivo de los	1 hora	Trabajo					1h			
3) Búsqueda de artículos y doc	1 hora	Trabajo					1h			
4) Lectura y comprensión de lc	2 horas	Trabajo					2h			
3) Búsqueda de artículos y doc	2 horas	Trabajo					2h			
4) Lectura y comprensión de lc	2 horas	Trabajo					2h			
6) Análisis visual de los datos (1 hora	Trabajo					1h			
7) Preprocesamiento de los da	0,5 horas	Trabajo					0,5h			
8) Visualización y estudio esta	1 hora	Trabajo					1h			
9) Análisis de estabilidad, tena	0,5 horas	Trabajo					0,5h			
10) Análisis de autocorrelaci	0,5 horas	Trabajo					0,5h			
11) Extracción de las zonas de	25 horas	Trabajo					13h	12h		
12) Implementación de distint	2 horas	Trabajo						2h		
13) Selección de las caracterís	2 horas	Trabajo						2h		
14) Construcción de las tablas	1 hora	Trabajo						1h		
15) Estudio prospectivo de los	3 horas	Trabajo						3h		
6) Análisis visual de los datos (1 hora	Trabajo						1h		
7) Preprocesamiento de los da	0,5 horas	Trabajo						0,5h		
8) Visualización y estudio esta	1 hora	Trabajo						1h		
9) Análisis de estabilidad, tena	0,5 horas	Trabajo						0,5h		
10) Análisis de autocorrelaci	0,5 horas	Trabajo						0,5h		
11) Extracción de las zonas de	25 horas	Trabajo						25h		
12) Implementación de distint	1 hora	Trabajo						1h		
13) Selección de las caracterís	2 horas	Trabajo						2h		
14) Construcción de las tablas	0,5 horas	Trabajo						0,5h		
15) Estudio prospectivo de los	1 hora	Trabajo						1h		
16) Desarrollo del diseño expe	5 horas	Trabajo						5h		
17) Pruebas de distintas alterr	1 hora	Trabajo						1h		
18) Análisis de los resultados, i	5 horas	Trabajo						5h		
16) Desarrollo del diseño expe	3 horas	Trabajo						3h		
17) Pruebas de distintas alterr	1 hora	Trabajo						1h		
18) Análisis de los resultados, i	10 horas	Trabajo						10h		
3) Búsqueda de artículos y doc	2 horas	Trabajo						2h		
4) Lectura y comprensión de a	4 horas	Trabajo						4h		
18) Análisis de los resultados, i	10 horas	Trabajo						10h		
19) Extracción de las zonas de	3 horas	Trabajo						3h		
19) Extracción de las zonas de	2 horas	Trabajo						2h		
20) Implementación de un sist	2 horas	Trabajo						2h		
20) Implementación de un sist	33 horas	Trabajo						3h	30h	
21) Realización de la memoria	45 horas	Trabajo							45h	
22) Puesta en marcha de la he	3 horas	Trabajo							3h	
23) Adquisición de más datos.	15 horas	Trabajo							15h	
24) Estudio de lo realizado par	4 horas	Trabajo							4h	
21) Completar la memoria con	5 horas	Trabajo							5h	
25) Realización de la presenta	12 horas	Trabajo							6h	6h

Figura 2.7: Uso de los recursos en la planificación completa.

7. Preprocesamiento de los datos en la pulsera de Microsoft.

Aquí todavía se estaba empezando y no se tenía un conocimiento completo de todo lo que había que hacer. Se continuó con la búsqueda de artículos (ítem 3) y su lectura (ítem 4), lo cual no fue del todo exitoso, ya que la mayoría de los artículos o usaban dispositivos ponibles no comerciales y/o situaban los dispositivos en otras partes del cuerpo: cintura, tobillo,... en lugar de utilizar la posición de los datos que nosotros teníamos disponibles: muñeca, o incluso utilizaban otro tipo de sensores como cámaras de vídeo para identificar la forma de andar de los usuarios, tal y como se explicará en la sección 3. Por tanto, se avanzó con la búsqueda y lectura de los artículos, pero se decidió posponer y continuar con ello más adelante, en las siguientes iteraciones, si fuera necesario. Respecto al análisis visual de los datos se realizó, pero los resultados eran extraños en algunas ocasiones, debido a que existían usuarios con tiempos de toma de datos negativos como se indicará en la subsección 4.2 y hubo que solucionarlos como se explica en el mismo apartado. En el preprocesamiento no hubo problemas, probándose diferentes opciones que se explicarán más adelante.

La fecha de comienzo fue el miércoles 6 de febrero y su finalización el miércoles 20, con la planificación cubierta.

2.5.3. Iteración 3

En esta tercera iteración se planificaron los ítems:

8. Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).
9. Análisis de estabilidad, tendencia, estacionalidad, etcétera de los datos.
10. Análisis de autocorrelaciones de los datos.
11. Extracción de las zonas de interés de los datos (de manera manual).

Ya se tenía un mayor conocimiento del problema y de los datos. Se visualizaron los datos en el dominio del tiempo, encontrando una librería de R llamada *dygraphs* [19] que facilitaba todo el proceso, permitiendo ampliar las zonas de interés y seleccionar los puntos exactos de cada uno de los datos. En el dominio de la frecuencia se trabajó con el Análisis de Fourier, pudiendo ver las frecuencias y amplitudes dominantes de cada uno de los usuarios. Se realizó el análisis de las autocorrelaciones, sin ningún problema, se incluirá en [5], así como la extracción de las zonas de interés que generó muchos problemas y dudas, porque ¿qué es realmente ruido? Existían frecuencias altas/bajas, pero a veces el comportamiento se repetía haciendo dudar entre si era ruido o un comportamiento característico del usuario.

La fecha de comienzo fue el miércoles 20 de febrero y su finalización el miércoles 6 de marzo, con la planificación cubierta a base de dedicar 30 horas más para invertirlas en finalizar la extracción de las zonas de interés de manera manual, que era mucho más costoso y difícil de lo planeado.

2.5.4. Iteración 4

En esta cuarta iteración se planificaron los ítems:

12. Implementación de distintas alternativas para la extracción de características a partir de los datos (dominio del tiempo y de la frecuencia).
13. Análisis estadístico de las distintas características extraídas para estudiar cuáles son más discriminantes (selección de las características de interés).
14. Construcción de las tablas de los datos con la información estructurada de entrada al clasificador.
15. Estudio prospectivo de los datos implementando algún clasificador y visualizando los resultados con las métricas necesarias.

A partir de la información extraída en la lectura de los artículos (ítems 3 y 4), se decidieron extraer características estadísticas tanto en el dominio del tiempo como en la amplitud obtenida con el Análisis de Fourier en el dominio de la frecuencia, junto con otras más que se explicarán en la sección 5.1. Después se construyeron las tablas de los datos donde cada fila se decidió que representará una ventana del usuario concreto en cada caso, teniendo cada usuario en cada toma de datos diversas ventanas de tamaño fijado. En este punto, se descubrió que el Análisis de Fourier no estaba dando resultados buenos, ya que al aplicar ventanas de suavizado se veía la existencia de una componente de continua que, tras investigar y dedicar bastante tiempo, se descubrió que se eliminaba restando la media. Respecto a la selección de características hubo muchos problemas por el diseño experimental utilizado, y se optó por hacer una selección a posteriori. Por último, se comenzó el estudio prospectivo aplicando k-vecinos más próximos y utilizando tanto el área bajo la curva (AUC) como la tasa de equierror (EER) como métrica para comparar los resultados.

La fecha de comienzo fue el miércoles 6 de marzo y su finalización el miércoles 20 de marzo, con la planificación cubierta a base de dedicar 15 horas más, aunque existía la necesidad de continuar con el ítem 15 y dejar de lado el 13, ya que las mejoras no eran significativas y aunque utilizar menos características reduciría el tiempo de cómputo, no se tenía esa necesidad y se podía seguir trabajando con todas las características.

2.5.5. Iteración 5

En esta quinta iteración se planificaron los ítems:

15. Continuación del estudio prospectivo de los datos implementando algún clasificador y visualizando los resultados con las métricas necesarias.
16. Desarrollo del diseño experimental del proyecto.

17. Análisis visual de los datos del giroscopio en la pulsera de Microsoft.
18. Preprocesamiento de los datos en la pulsera de Microsoft.

Se continuó con el estudio prospectivo, ya que parecía que no se estaba utilizando la característica temporal de la serie y se decidió probar si el módulo de los datos, que se obtiene con la fórmula (2.1),

$$\sqrt{X^2 + Y^2 + Z^2} \quad (2.1)$$

proporcionaba los mismos, mejores o peores resultados. También se probaron más técnicas de selección de características. El objetivo era ir tomando decisiones y viendo qué es lo que realmente merecía la pena aplicar a los datos, tanto respecto a preprocesado como a algoritmos de clasificación.

Para aplicar más clasificadores, se decidió cambiar el diseño experimental del proyecto (ítem 16) y pasar de trabajar con datos de la misma clase en una especie de aprendizaje no supervisado, a tener una clase binaria: usuario auténtico/usuario impostor con aprendizaje supervisado.

Al aplicar más clasificadores se descubrió un comportamiento extraño en los resultados, que parecía debido a la existencia de sobreaprendizaje en los datos. Esto paralizó el resto del trabajo para buscar soluciones a este problema. Se barajó incluso la posibilidad de recoger nuevos datos en la siguiente iteración.

La fecha de comienzo fue el miércoles 20 de marzo y su finalización el miércoles 3 de abril, con únicamente cubiertos los ítems 13, 15 y 16, pero incluso de manera incompleta por la presencia de problemas, por el momento, desconocidos.

2.5.6. Iteración 6

En esta sexta iteración se planificaron los ítems:

15. Continuación del estudio prospectivo de los datos implementando más clasificadores y visualizando los resultados con las métricas necesarias.
57. Puesta en marcha de la herramienta de adquisición realizada en TFGs anteriores.
58. Adquisición de más datos.
17. Análisis visual de los datos del giroscopio en la pulsera de Microsoft.
18. Preprocesamiento de los datos en la pulsera de Microsoft.

Se dio prioridad a solucionar los problemas con los clasificadores, ya que continuar no servía de nada. Se cambió el entorno de programación de R a Python por la sospecha de que las librerías de R utilizarán algún tipo de ventana de suavizado que se aplicará a los resultados, generando ese

sobreaprendizaje. También se utilizó Weka como herramienta para aplicar clasificadores y ver la diferencia de resultados entre R y Weka. No obstante, ese no era el problema.

Por otro lado, se decidió cambiar el rumbo del proyecto y poner en marcha la aplicación de recogida de datos para conseguir nuevos y más cantidad de usuarios. La aplicación recogía datos desde 2 dispositivos: una pulsera de *Microsoft* y un reloj de *Motorola*. La pulsera funcionó rápidamente, sin problemas, pero el reloj no se consiguió que funcionara. Se probó desde diferentes teléfonos móviles, el primero con Android 8.1.0 y el segundo con Android 6.0 que era la misma versión que se había utilizado cuando se recogieron los datos con los que se estaba trabajando, pero no se consiguió poner en funcionamiento. La falta de conocimientos de Android impidió continuar por este camino.

Se hicieron muchas pruebas, entre ellas, implementar k-vecinos más próximos como clasificador manualmente, paso a paso, sin utilizar ninguna librería, pero los resultados coincidían. Después, se probó a eliminar características, resultando que posiblemente ahí sí estuviera el problema. Por lo que se invirtió tiempo en investigar y tratar de solucionar los problemas, buscando conseguir un buen estudio prospectivo (ítem 15), ya que después, el trabajo con el resto de los dispositivos iría más rápido.

La fecha de comienzo fue el miércoles 3 de abril y su finalización el 17 de abril, con únicamente el ítem 15 cubierto a medias y 10 horas más dedicadas.

2.5.7. Iteración 7

En esta séptima iteración se planificaron los mismos ítems que en la anterior:

15. Continuación del estudio prospectivo de los datos implementando más clasificadores y visualizando los resultados con las métricas necesarias.
57. Puesta en marcha de la herramienta de adquisición realizada en TFGs anteriores.
58. Adquisición de más datos.
17. Análisis visual de los datos del giroscopio en la pulsera de Microsoft.
18. Preprocesamiento de los datos en la pulsera de Microsoft.

Respecto al estudio prospectivo y tras solucionar el problema del sobreaprendizaje se decidió aprovechar la debilidad que se tenía, que era tener pocos usuarios, para hacer un estudio básico de la biometría en profundidad, ya que actualmente, en los artículos, no estaba disponible para dispositivos comerciales como los que se estaban utilizando. Esto conllevaba la realización de las siguientes tareas:

- Estudiar y realizar tablas para cada usuario y no en media como se hacía hasta ahora.

- Centrarse en un único clasificador y estudiarlo en profundidad. El elegido fue k-vecinos más próximos por su simplicidad e introducir pocos parámetros. Entre las pruebas se encontrarían:
 - Aplicar diferentes tamaños de ventana para ver si a mayor tamaño, mejores resultados.
 - Aplicar filtros de media móvil para ver si beneficia a los resultados.
 - Probar a eliminar, sobre los datos a los que ya se les ha eliminado el ruido de manera manual, aquellas ventanas de cada usuario que no pasaran de un determinado valor umbral de autocorrelación fijado de antemano.
 - Probar a realizar el estudio con los datos enteros sin eliminar el ruido, eliminando ventanas de baja autocorrelación, para ver si se podía utilizar como técnica de aquí en adelante, como solución al ítem con ID 54 de la lista de tareas de la tabla 2.3.

En esta iteración se cumplieron parte de las tareas añadidas mencionadas continuando con el ítem 15 del estudio prospectivo.

También se dedicó tiempo al ítem 57 y a solucionar los problemas de la aplicación para la adquisición de datos con el reloj Motorola, buscando ayuda entre los profesores con conocimientos de Android, pero no se consiguió nada.

La fecha de comienzo fue el miércoles 17 de abril y su finalización el miércoles 1 de mayo empleando 15 horas más, invertidas en solucionar el problema de la aplicación Android y a tomar todas las decisiones para centrar el proyecto por los nuevos caminos que estaba tomando.

2.5.8. Iteración 8

En la octava iteración se decidió dar más prioridad a poner en marcha la herramienta de adquisición de datos. La razón era que ya se habían dedicado muchas horas y se estaba cada vez más cerca, aunque por razones desconocidas y la falta de conocimiento de Android no se terminaba de conseguir. Los ítems planificados fueron:

15. Continuación del estudio prospectivo de los datos con las nuevas tareas fijadas en la iteración anterior.
57. Puesta en marcha de la herramienta de adquisición realizada en TFGs anteriores.
58. Adquisición de más datos.

Tras muchas pruebas e invertir mucho tiempo, cuando se estaba terminando la iteración, se consiguieron resolver todos los problemas de la aplicación Android que permitía la adquisición de nuevos datos. En el teléfono móvil que se estaba utilizando, ya funcionaban ambos dispositivos, la pulsera de *Microsoft* y el reloj *Motorola*. No obstante, había un problema al recoger el identificador

único del teléfono móvil (*International Mobile Equipment Identity o IMEI*), porque no lo reconocía ni lo guardaba y hacía que todos los campos en la Base de Datos aparecieran en posiciones erróneas generando problemas entre los campos numéricos y carácter. Dadas las fechas en las que nos encontrábamos y todo lo que habíamos hecho, se decidió dejar apartada la adquisición de datos y continuar con ello, tras la finalización del TFG, ya que como se formaba parte de un proyecto de investigación, tanto la autora del proyecto como los profesores estaban dispuestos a continuar una vez terminado. De esta manera, el funcionamiento de la aplicación no era tiempo perdido.

Respecto al estudio prospectivo, haciendo pruebas iniciales buscando qué más cosas se podían probar, se descubrió que modificando los scores obtenidos con la técnica de clasificación mediante una fusión fijado el tamaño de la ventana y el solapamiento de la misma empleando diversas métricas como la media, la mediana, el mínimo o el máximo mejoraba los resultados.

La fecha de comienzo fue el miércoles 1 de mayo y su finalización el miércoles 15 de mayo tras emplear 15 horas más.

2.5.9. Iteración 9

En esta novena iteración se decidió profundizar en el estudio prospectivo de la pulsera *Microsoft* y el sensor *Acelerómetro*, tomando las decisiones finales que generasen un sistema de reconocimiento con el que probar los diferentes escenarios experimentales desarrollados en el diseño experimental creado. Esto sería el ítem 51 de la tabla 2.3. Respecto al estudio prospectivo se fijaron las siguientes tareas.

- Aplicar diferentes tamaños de ventana para ver si un tamaño intermedio genera mejores resultados que uno más pequeño o más grande.
- Aplicar filtros de la media móvil para ver si beneficiaba a los resultados.
- Probar el efecto del ruido, eliminando las ventanas de baja autocorrelación, sobre los datos donde ya se ha eliminado el ruido de manera manual.
- Con los datos originales, probar a eliminar las ventanas de baja autocorrelación.
- Aplicar varias ventanas sobre los scores producidos con *k-vecinos más próximos* utilizando distintos estadísticos: media, mediana, mínimo, máximo... y distintos tamaños de ventana y solapamiento de la fusión.

Esto implicaba realizar las siguientes tareas (se utiliza el mismo identificador que en la tabla 2.3).

15. Completar el estudio prospectivo de los datos con la pulsera de *Microsoft* y el sensor acelerómetro realizando las nuevas tareas que se acaban de mencionar.
51. Tomar las decisiones necesarias para generar un buen sistema de reconocimiento final.

54. Conseguir un sistema automático para extraer las zonas de interés de los datos.
15. Aplicación de las decisiones tomadas en todas las posibilidades del entorno experimental para la pulsera de *Microsoft* y el acelerómetro.

Estas tareas darían por finalizado el estudio con la pulsera de *Microsoft* y el sensor acelerómetro. Dadas las fechas en las que nos encontrábamos, finalizando las horas totales que se tenían que dedicar al proyecto, se decidió reducir lo que se quería hacer. De manera que para el otro sensor, giroscopio de la pulsera de *Microsoft* y ambos sensores, acelerómetro y giroscopio del reloj *Motorola* se haría lo siguiente:

- Automatizar la extracción de las zonas de interés de los datos. Esto eliminaba los ítems 22, 34 y 44, completando el 54.
- Estudiar las autocorrelaciones y si resultaban parecidas a las ya estudiadas en la pulsera *Microsoft* con el sensor acelerómetro, utilizar tanto en el sensor giroscopio como en el otro dispositivo (*Motorola*) las mismas conclusiones de preprocesamiento. Bajo el argumento de que las autocorrelaciones y su visualización eran similares. Esto eliminaba los ítems 18, 30 y 40, utilizando en su lugar el ítem 7.
- Dado que la extracción de características (ítem 12) se había hecho utilizando como referencia los artículos leídos, utilizar las mismas en ambos sensores y dispositivos, eliminado los ítems 23, 35 y 45, gracias a ser sustituidos por el 12.
- De la misma manera que el preprocesamiento y la extracción de características, reutilizar el estudio prospectivo, eliminado los ítems 26, 38 y 48, sustituyéndolos por el 15.
- Realizar la selección de características una única vez sobre el sistema de reconocimiento final en un único dispositivo y sensor que sería: *Microsoft acelerómetro*. Con el objetivo de ver si esto era una buena manera de mejorar los resultados en esta biometría. Eliminando los ítems 24, 36 y 46 para realizar sólo el 13.

Lo que sí se seguiría haciendo serían las siguientes tareas:

- Para **Microsoft Giroscopio**:
 17. Análisis visual de los datos.
 19. Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).
 21. Análisis de autocorrelaciones de los datos.
 25. Construcción de las tablas de los datos con la información estructura de entrada al clasificador.
- Para **Motorola Acelerómetro**:

29. Análisis visual de los datos.
 31. Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).
 33. Análisis de autocorrelaciones de los datos.
 37. Construcción de las tablas de los datos con la información estructura de entrada al clasificador.
- Para **Motorola Giroscopio**:
39. Análisis visual de los datos.
 41. Visualización y estudio estadístico de los datos del corpus (dominio del tiempo y de la frecuencia).
 43. Análisis de autocorrelaciones de los datos.
 47. Construcción de las tablas de los datos con la información estructura de entrada al clasificador.

La fecha de comienzo fue el miércoles 15 de mayo y su finalización el miércoles 29 de mayo habiendo completado todas las tareas de la pulsera *Microsoft* con el sensor acelerómetro, incluyendo la construcción del sistema de reconocimiento final (ítem 51), la automatización para la extracción de las zonas de interés (ítem 54) y el estudio prospectivo en todas las posibilidades del entorno experimental (ítem 15). Por otro lado, también se realizó la visualización de los datos (ítems 17, 19, 29, 31, 39 y 41), el estudio de las autocorrelaciones (ítems 21, 33 y 43) y la construcción de las tablas de los datos (ítems 25, 37 y 47), correspondientes a los sensores y dispositivos restantes. En esta iteración se han dedicado 30 horas más, ya que en el momento de finalización de las tareas correspondientes al dispositivo *Microsoft* y el acelerómetro se invirtieron escribiendo la memoria final del TFG.

2.5.10. Iteración 10

En esta décima iteración ya se tenía muy claro todo lo que se iba a hacer para dar por finalizado el proyecto. Como se tenían limpios los datos y construidas las tablas de características en ambos sensores y dispositivos, se planificaron las siguientes tareas.

1. Aplicar el sistema de reconocimiento final construido (ítem 51) en todas las posibilidades del entorno experimental para la pulsera de *Microsoft* y el giroscopio.
2. Aplicar el sistema de reconocimiento final construido (ítem 51) en todas las posibilidades del entorno experimental para el reloj *Motorola* y el acelerómetro.
3. Aplicar el sistema de reconocimiento final construido (ítem 51) en todas las posibilidades del entorno experimental para el reloj *Motorola* y el giroscopio.

4. (Ítem 13) Análisis estadístico de las distintas características extraídas para estudiar cuáles son más discriminantes (selección de las características de interés) con la pulsera de *Microsoft* y el acelerómetro.
5. (Ítem 56) Realización de la memoria.

Como era de esperar, al tener las cosas claras de lo que había que hacer y el código construido de cuando se había utilizado para la pulsera de *Microsoft* y el acelerómetro, no surgieron problemas en las 3 primeras tareas. En la selección de características, cuarta tarea planificada en esta iteración, surgieron dos maneras de actuar y se decidió probar ambas, como se explicará en el capítulo de selección de características de [5]. La fecha de comienzo fue el miércoles 29 de mayo y su fecha de finalización el miércoles 12 de junio, habiendo realizado todas las tareas pensadas para el TFG. Ahora sólo faltaba terminar de realizar la memoria (ítem 56), la cual se había avanzado mucho, tras dedicar 30 horas más.

2.5.11. Iteración 11

Esta onceava iteración es añadida a la planificación inicial que se tenía, donde se había pensado finalizar el día 3 de junio, dentro del periodo de tiempo de la iteración 10. No obstante, ya se había terminado todo el trabajo y sólo faltaba terminar de escribir la memoria, por lo que se planificaron las siguientes tareas de la tabla 2.3.

56. Realización de la memoria. Incluye versiones y revisiones.
61. Realización de la presentación para la defensa oral.

La fecha de comienzo fue el miércoles 12 de junio y su fecha de finalización el miércoles 26 de junio, con todo el trabajo realizado y las memorias presentados en la solicitud de defensa de la convocatoria ordinaria de la *Universidad de Valladolid*. Para la iteración completa se realizaron un total de 30 horas comprendidas entre las fechas indicadas.

Una vez terminadas todas las iteraciones, se muestra en la figura 2.8 el calendario con la fecha de inicio y fin reales (color naranja), los reuniones no realizadas (color amarillo con una X) y las reuniones realizadas a mayores (color rojo). Al principio se necesitaron más reuniones para ubicarse en el problema y conseguir una extracción de las zonas de interés suficientemente buena. En marzo también se realizaron más reuniones por la falta de conocimientos del procedimiento experimental seguido en biometría. En abril se empezaron a probar más clasificadores y existió el problema del sobreaprendizaje, comenzando a poner en funcionamiento la aplicación de captura de datos, lo que generó muchas más reuniones. En mayo, el tiempo se acababa y hubo que hacer reuniones para ubicar el trabajo que se incluía en la presente memoria. Mientras que en junio se realizaron reuniones para mostrar a los tutores el resultado final, realizando en paralelo la memoria, que llevo bastante más tiempo que el planificado, terminando el proyecto 20 días más

tarde, pero a tiempo aún de la convocatoria ordinaria. Se deja una reunión pendiente después de presentar la solicitud de defensa para enseñar a los tutores la presentación PowerPoint que se utilizará en el tribunal. En la parte inferior se pueden ver los días totales que se han dedicado cada mes.

2.6. Proceso técnico completo

En esta subsección se muestran todas las herramientas y librerías más relevantes que se irán utilizando a lo largo del desarrollo del proyecto, tanto las planificadas como las finalmente utilizadas.

Como herramientas para la comunicación alumno-profesores:

- UVa Webmail alumnos: fuente de comunicación principal, enviando correos para concretar reuniones, dudas y problemas.
- Life Size Cloud: realización de videoconferencias durante el periodo de vacaciones de Semana Santa por la necesidad de tomar decisiones [20].
- WeTransfer: herramienta para enviar archivos con una capacidad máxima permitida de 2GB [21].

Para la realización de informes, tablas y resúmenes:

- Overleaf: edición de la memoria final en \LaTeX .
- Excel: realización de tablas resumen de los datos, para mostrar toda la información agrupada de manera más clara y sencilla.
- Word: realización de documentos auxiliares rápidos y para tomar notas.

Para la edición de las imágenes incluidas en el proyecto:

- PhotoJoiner: unión de múltiples fotos juntas [22].
- Fotor: edición de fotos, añadir texto [23]...

Como lenguajes de programación:

- R: mayoritariamente, trabajando con RStudio [24] como editor de código, debugging y herramienta de visualización.
- Python: utilizando Anaconda [25] como plataforma de programación.

Entre las librerías de R más relevantes:

2019

enero		febrero		marzo		abril		mayo		junio		julio	
jueves		01		01	REUNION	01	REUNION	01	FIESTA	01	REUNION	01	
viernes		02		02		02	REUNION	02		02		02	
sábado		03		03		03		03		03		03	
domingo		04	REUNION	04		04		04		04	REUNION	04	REUNION
lunes		05		05		05		05		05		05	
martes	01	06	X	06	REUNION	06	REUNION	06		06		06	
miércoles	02	07	REUNION	07		07		07		07		07	REUNION
jueves	03	08	REUNION	08		08		08		08		08	
viernes	04	09		09		09		09		09		09	
sábado	05	10		10		10		10		10		10	
domingo	06	11		11		11		11		11		11	
lunes	07	12		12		12		12		12		12	
martes	08	13		13	REUNION	13	REUNION	13		13		13	
miércoles	09	14	REUNION	14		14		14		14		14	
jueves	10	15		15	REUNION	15		15		15		15	
viernes	11	16		16		16		16		16		16	
sábado	12	17		17		17		17		17		17	
domingo	13	18		18		18		18		18		18	
lunes	14	19		19		19		19		19		19	
martes	15	20	REUNION	20		20		20		20		20	
miércoles	16	21		21	REUNION	21	SEMANA SANTA	21		21		21	
jueves	17	22		22		22		22		22		22	
viernes	18	23		23		23		23		23		23	
sábado	19	24		24		24		24		24		24	
domingo	20	25		25		25		25		25		25	
lunes	21	26		26		26		26		26		26	
martes	22	27	X	27	REUNION	27	REUNION	27	REUNION	27		27	
miércoles	23	28		28		28		28		28		28	
jueves	24			29		29		29		29		29	
viernes	25			30		30		30		30		30	
sábado	26			31		31		31		31		31	
domingo	27												
lunes	28												
martes	29												
miércoles	30	REUNION											
jueves	31												
viernes													
		8 días		26 días		26 días		27 días		22 días			

Figura 2.8: Calendario real del desarrollo del proyecto.

- xlsx: paquete para leer, escribir y formatear archivos de Excel.
- Dygraphs: librería muy utilizada y realmente útil. Es una interfaz R para la biblioteca de gráficos de JavaScript de dygraphs. Proporciona facilidades para trazar datos de series temporales. Incluye [19]:
 - Trazar automáticamente la serie temporal.
 - Visualizar la serie y los ejes, los cuales son fácilmente configurables.
 - Funciones interactivas que permiten zoom y recorte de manera rápida y fácil, así como la señalización exacta de los puntos junto con sus valores.
 - Permite mostrar barras superiores/inferiores (intervalos de predicción) alrededor de la serie.
 - Superposición de gráficos que incluyen regiones sombreadas, líneas, puntos anotados.
 - Se incrusta sin problemas en documentos RMarkdown y aplicaciones web Shiny.
- intervals: herramienta para comparar conjuntos de puntos e intervalos. Utilizada para realizar la intersección de diversos intervalos.
- moments: herramienta para calcular medidas estadísticas de los datos como momentos, acumulados, coeficiente de asimetría, curtosis o test relativos.
- DescTools: herramienta para el cálculo de medidas estadísticas descriptivas.
- FNN: librería muy utilizada para el estudio prospectivo de los datos. Implementa los algoritmos de clasificación y regresión de búsqueda del vecino más cercano, KD-tree, cover-tree.
- PRROC: librería para el cálculo de las medidas relacionadas con las curvas ROC como el área bajo la curva (AUC).
- devtools: herramienta para facilitar el desarrollo de paquetes en R. Se utiliza para descargar la librería *ROC* de Github a través de: `install_github("davidavdav/ROC")` que permite el cálculo de la tasa de equierror (EER) a partir de la curva ROC.
- ggplot2: librería para mejorar de manera fácil la representación gráfica.
- reshape2: permite la reestructura y agregación de datos de manera flexible.
- stats: utilizada para representar test estadísticos basados en la búsqueda de homocedasticidad en los datos (igualdad de varianzas) con la orden *fligner.test* y test no paramétrico de Mann-Whitney-Wilcoxon con la orden *wilcox.test*.
- FSelector: contiene funciones para la selección de atributos de un conjunto de datos determinado. La selección de subconjuntos de atributos es el proceso de identificar y eliminar la mayor cantidad de información irrelevante y redundante que sea posible. Se ha utilizado la

orden *relief* que es un algoritmo que encuentra los pesos de atributos continuos y discretos basándose en la distancia entre las instancias.

Las librerías de Python utilizadas han sido:

- Pandas: herramienta de manipulación de datos de alto nivel desarrollada por Wes McKinney. Está construida con el paquete Numpy y su estructura de datos clave es el DataFrame, el cuál te permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables [26].
- Sklearn: biblioteca para el aprendizaje automático de software libre. Simple y eficiente para análisis de datos y Data Mining. Incluye algoritmos de clasificación, regresión y análisis de grupo entre los cuales están máquinas de vectores soporte, bosques aleatorios, Gradient Boosting, K-medias y DBSCAN. También permite técnicas de reducción de la dimensionalidad, selección de modelos y preprocesamiento. Diseñada para interoperar con las bibliotecas numéricas y científicas NumPy y SciPy [27].
- Matplotlib: biblioteca para la generación de gráficos a partir de los datos contenidos en listas o arrays. Diseñada para interoperar con la biblioteca numérica NumPy [28].

Capítulo 3

Estado del Arte

A lo largo de este capítulo se va a exponer la situación actual de la biometría, centrándonos en el estudio de aquellas cosas que nos afectan, como la manera de actuar con los datos o las técnicas de aprendizaje automático que se emplean, haciendo un resumen de los resultados obtenidos en investigaciones similares utilizando dispositivos ponibles y el comportamiento de las personas como patrón.

El reconocimiento biométrico consiste en aplicar técnicas estadísticas y matemáticas sobre las características fisiológicas o del comportamiento de un individuo para su reconocimiento, ya sea identificación o verificación. Estos sistemas, como se ha dicho en el capítulo 1 de la Introducción, presentan una serie de ventajas, tales como que los usuarios no necesitan recordar claves complejas para su autenticación ni llevar consigo llaves, tarjetas u otros objetos físicos, que pueden perderse o transferirse.

Existen sistemas biométricos tradicionales y portátiles. Los sistemas portátiles, por su naturaleza, están siempre con el usuario pudiendo almacenar los datos dentro del dispositivo, siendo capaces de leer la señal del sujeto en cualquier momento y por tanto, permitiendo la autenticación continua, mientras que los sistemas biométricos tradicionales son generalmente colocados en un lugar fijo, menos susceptibles de deteriorarse así como más fácilmente reemplazables, haciendo uso de procesos más costosos computacionalmente, ya que pueden utilizar fuentes externas de energía [6]. Ejemplo de sistema biométrico tradicional es una característica de Windows 10 llamada *Windows Hello* [29] que permite al usuario autenticarse usando la cara, el iris o la huella digital.

La creciente popularidad de los dispositivos portátiles está llevando a nuevas formas de interactuar con otros dispositivos inteligentes y con otras personas. Los *wearables* equipados con una serie de sensores son capaces de capturar los rasgos fisiológicos y de comportamiento del propietario, resultando apropiados para biometría, siendo éstos los que se van a utilizar en el presente proyecto. Los sensores predominantes en los dispositivos portátiles actuales son [6]:

- **Sensores de luz:** dependiendo de la resolución del sensor, pueden ser utilizados para medir la intensidad de la luz, como por ejemplo los sensores fotopletismográficos (PPG) [30]

que miden el volumen de cambio sanguíneo dentro del tejido microvascular, o proporcionar imágenes completas, como es el caso de los lectores de huellas dactilares [31] o cámaras digitales [32] que pueden capturar las características fisiológicas como la cara u otras características corporales como la forma de andar de los individuos a través de lo que se llaman *técnicas de visión*.

- **Sensores de fuerza:** mide la fuerza que afecta al dispositivo de medición, ya sea originada por el movimiento, ejemplo de ello es el acelerómetro tridimensional [33, 34] o por la fuerza de Coriolis como hace el giroscopio o el campo magnético de la Tierra con el magnetómetro o la presión del aire con el barómetro.
- **Sensores eléctricos:** mide la actividad eléctrica de algunas partes del cuerpo, como por ejemplo, un electrocardiograma para el corazón [35] o cómo cambia una corriente cuando se aplica al cuerpo, como por ejemplo, la conductividad de la piel con un sensor de respuesta galvánica de la piel [36].
- **Sensores de temperatura:** funcionan como una cámara infrarroja. Se captura la energía infrarroja y se transforma en una señal digital que representa la temperatura. Los sensores de temperatura de la piel generalmente se colocan a una distancia muy corta o en contacto directo con la piel. La miniaturización de la tecnología ha permitido el desarrollo de pequeños sensores de temperatura de la piel que pueden incorporarse en casi cualquier dispositivo electrónico, como los dispositivos ponibles [37].
- **Sensores de sonido:** un micrófono traduce las ondas de sonido que viajan por el aire en una señal eléctrica. Hay micrófonos comerciales que están preparados para capturar la voz humana a una distancia razonable (60dB a 1 metro), ya que la voz de una persona se define por las características fisiológicas del sistema respiratorio de la persona [38].
- **Sensores de localización:** El Sistema de Posicionamiento Global (GPS) consta de 32 satélites y cualquier número de receptores GPS ubicados en la superficie de la Tierra. Un receptor GPS utiliza la señal de cuatro satélites de línea de visión diferentes para triangular la ubicación del dispositivo, ofreciendo sus coordenadas (longitud y latitud), proporcionando información de comportamiento solo con respecto a la ubicación del sujeto.

En este trabajo, como se explicará en el apartado 4.1, se va a trabajar con sensores portátiles de fuerza. Pero la biometría es un problema difícil en continuo estudio, con cada vez más tipos de sensores diferentes, que algún día podrán ser usados de forma complementaria para conseguir mejores resultados.

Una vez adquirida una muestra de datos del usuario mediante el sensor existen dos formas de abordar el trabajo: considerando toda la muestra adquirida o dividiendo esa muestra en marcos temporales. Los artículos encontrados trabajan de la segunda forma ya que justifican que de esta manera se captura la variabilidad del individuo con el tiempo, pero dependiendo de cuál lo hace de diferente forma. Todos ellos consideran ciclos de marcha y que la forma de caminar humana

es un movimiento periódico, compuesto por un paso de la pierna derecha y un paso de la pierna izquierda. Es decir, un ciclo de marcha empieza cuando un pie toca el suelo y termina cuando el mismo pie toca el suelo nuevamente como se muestra en la figura 3.1. Dentro de los trabajos leídos cabe destacar [39] por utilizar solamente la dimensión Z del acelerómetro para hacer la partición del ciclo de la forma de andar, ya que afirma existir una asociación entre la fuerza de reacción del suelo y la fuerza de la señal de este eje, que forma picos de gran magnitud y busca esos cambios del eje Z para dividir la señal en ventanas. Otros como [40] utilizan el periodo de la señal para detectar los ciclos y hacer la división. Por último, en [41] se hace una revisión extensa del enfoque de ventanas, mostrando el tamaño utilizado, en segundos, de distintas publicaciones en las que se realizan diversas actividades, no solo la de caminar y se sitúan distinto número de acelerómetros en distintas posiciones, que también se indica. Considera la creación de ventanas, para cada actividad, en función del flujo de datos del sensor y los cambios que se producen, pudiendo identificar dichos cambios a través de un análisis de variaciones en las características de la frecuencia de la señal; o bien detectando el contacto inicial y final del pie con el suelo a través de la aceleración lineal del pie. Introduce la superposición entre ventanas adyacentes, lo que llamaremos “solapamiento” y demuestra que su efecto es beneficioso para el reconocimiento de actividades periódicas como caminar o correr, y estáticas como estar de pie o sentado, pero de utilidad cuestionable para la detección de actividades esporádicas, en las que su naturaleza es más compleja e intercalada. La publicación [42] considera ventanas con 20 % de solapamiento y [43, 44] consideran un 50 %.

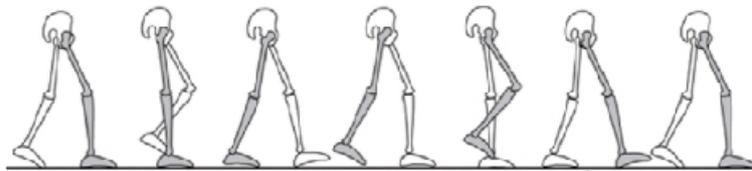


Figura 3.1: Esquema de un *ciclo de marcha*.

Por otro lado, casi nunca se utiliza la señal cruda de los datos, ya que un buen preprocesamiento puede ayudar a mejorar los resultados. Lo que todos los artículos hacen es eliminar el ruido, destacando [40, 45–48] por hacerlo asignando pesos a los datos a través del filtro Weighted Moving Average (WMA), en [46, 47] también se eliminan los falsos mínimos a través del ciclo medio, calculando aquellos puntos fuera del rango ($media \pm desviacion_estandar$) o bien con filtros de la mediana como en [49, 50] o filtros *lowpass* o *highpass* para eliminar las interferencias fuera de la banda como hacen [39, 45, 51]. La segunda técnica más aplicada es la de la interpolación por tener los datos disponibles en intervalos de tiempo irregulares, [46, 47, 52] aplican una interpolación de spline, mientras que [39, 40, 45] justifican que utilizar una interpolación lineal es suficiente y más sencillo. Por último, la mayoría de los estudios analizados normalizan los datos, tanto si trabajan en el Dominio del Tiempo como si lo hacen con las amplitudes de Fourier en el Dominio de la Frecuencia, destacando [39, 46, 47, 51, 52], pero ninguno de ellos compara el efecto de lo que ocurre si no se normalizan los datos.

Una vez se ha decidido si trabajar con toda la muestra o con una división de ella en ventanas y el

preprocesado que aplicar a los datos, hay que decidir si trabajar con la señal preprocesada cruda o si realizar una extracción de características que represente al usuario. Cuando se trabaja con la señal preprocesada cruda se suele aplicar el método de Dynamic Time Warping (DTW) para obtener la distancia entre las ventanas o las señales, como ocurre en [6]. Pero la mayoría de las investigaciones se centran en la extracción de características, utilizando tanto el Dominio del Tiempo como el Dominio de la Frecuencia. En el Dominio de la Frecuencia, lo más utilizado es la transformada de Fourier, donde destacan las publicaciones [40, 45, 49, 51, 53]. No obstante también se aplica la Transformada del Coseno Discreta (DCT) en [6, 34, 40] donde se utilizan sus coeficientes para intentar representar al usuario. En el Dominio del Tiempo lo más utilizado son medidas estadísticas como la media y la mediana en [46, 47, 54, 55], la desviación estándar en [39, 55, 56], el mínimo y el máximo en [34, 45, 55, 57, 58]. También se extraen otras características como la curtosis en [34, 54], el coeficiente de asimetría en [53] o los ratios medios de las componentes XY, XZ o YZ como representantes de la gravedad, en [45]. Otro tipo de características incluidas son las correlaciones en [40, 51, 52] o la información sobre los ángulos de los ejes en [57]. Respecto a las características, cabe destacar [57], donde se demuestra como la precisión mejora cuando se agregan valores estadísticos adicionales al vector de características y se indica que el vector óptimo es aquel que contiene tantos pocos datos como sea posible sin perder ningún criterio de información discriminativo. Entre las técnicas para reducir la dimensión del vector de características destacan Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA) y Support Vector Machines (SVM), por aplicarse en [49], mientras que PCA también se aplica en [39].

Se ha hecho un resumen de varios artículos en los que se han utilizado dispositivos wearables, para los que se va a indicar la siguiente información. En todos los casos se está utilizando el sensor de un teléfono móvil (smartphone).

- **Técnica ML:** Técnica de aprendizaje automático utilizada.
- Referencia al **artículo**, junto con el nombre del autor y el año.
- Tipo de **sensor** utilizado entre los explicados anteriormente y que nos interesan, que son los de fuerza.
- **Modo** de reconocimiento biométrico: identificación (I) o verificación (V).
- Los resultados, indicando la tasa de equierror (**EER**), la precisión (**H**) y/o la tasa de falsos positivos (**FPR**), según qué información esté disponible.
- Número de **sujetos** en la Base de Datos.
- El número de **características** utilizadas.

Los algoritmos de Machine Learning usados en biometría intentan resolver dos problemas diferentes.

- **Identificación biométrica:** modo I, es un problema de clasificación multiclase.
- **Verificación biométrica:** modo V, es un problema de clasificación de una clase. Más frecuente en los sistemas portátiles.

La salida de los algoritmos ejecutados es un valor numérico que mide el grado de similitud entre la señal consultada y un sujeto registrado. Después de obtener este resultado, generalmente se aplica un umbral t para determinar la decisión final. La variación de t ajusta las tasas de falsos positivos y falsos negativos (FPR y FNR, respectivamente), generando lo que se llama las curvas Receiver Operating Characteristic (ROC). Las métricas más usadas en biometría y que se van a utilizar en el estudio de los diversos artículos son además de la tasa FPR, la tasa de equierror (EER) que es el punto de la curva ROC en el que FPR es igual a FNR, y una medida más general de la precisión (H) que corresponde con el número de veces que el sistema produce la decisión correcta.

Hay muchas técnicas pero las más frecuentes para este tipo de problemas, y cuyos resultados aparecen en la tabla 3.1 son:

- **K -Nearest Neighbour (KNN):** o k -vecinos más próximos es un método de aprendizaje perezoso por almacenar vectores de características en el conjunto de entrenamiento y retrasar todo el procesamiento hasta la clasificación. Muy utilizado y popular por su simplicidad y efectividad.
- **Support Vector Machines (SVM):** o máquinas de vectores soporte es un método estadístico que construye un hiperplano que separa de manera óptima las diferentes clases de las muestras de entrenamiento. La efectividad de SVM depende del kernel seleccionado y de un parámetro de margen que describe la influencia de una sola muestra en el hiperplano. En [56, 59] los investigadores utilizaron el acelerómetro para verificar la identidad de los sujetos mientras realizaban gestos, pudiendo verificar la identidad del sujeto solo mientras caminaban en un ambiente muy restringido. En cambio, en [60] propusieron un sistema multimodal que consistía en acelerómetro, giroscopio y posicionamiento GPS para verificar la identidad de un sujeto, consiguiendo resultados prometedores, pero utilizando únicamente 3 personas. Por otro lado, destaca [61] por obtener el 100% de precisión utilizando los datos del acelerómetro, pero su estrategia de clasificación no fue exactamente un sistema biométrico, sino una mezcla de identificación y verificación, con una población también muy limitada de 5 individuos.
- **Gaussian Mixture Model (GMM):** o modelo de Mixtura Gaussiana, es un modelo probabilístico que asume que todas las muestras del mismo sujeto pueden generarse por una suma ponderada de un número finito de distribuciones gaussianas. Los pesos de cada distribución y sus parámetros se obtienen a través de diferentes métodos de ajuste, por ejemplo, el más común en la literatura es el de maximización de las expectativas (EM). En un sistema de verificación, se debe establecer un umbral de probabilidad para seleccionar muestras como válidas para ese GMM. En un sistema de identificación biométrica, la muestra de consulta

se pasa a través de todos los GMM de los sujetos, y se selecciona el que tiene más probabilidad. Pero es un método más utilizado para los sonidos emitidos por el cuerpo, como por ejemplo, del corazón, donde los experimentos [62, 63] han logrado precisiones entre 0.86 y 1 con poblaciones de sujetos de tamaño medio (entre 10 y 80 individuos). También se ha empleado en verificación utilizando el acelerómetro y la respuesta galvánica de la piel [33, 34] con peores resultados de EER y FPR por encima de 0.14 en todos los casos.

- **Hidden Markov Model (HMM):** o modelo oculto de Markov es un tipo particular de red Bayesiana, donde el sistema realiza la transición de un estado a otro según las observaciones y un conjunto de probabilidades de transición que se desconocen previamente. Los HMM se han utilizado ampliamente en varios problemas de aprendizaje automático, pero son especialmente conocidos por sus aplicaciones en reconocimiento de voz [64], donde se ha logrado EERs promedio de 0.10 con 48 individuos diferentes.
- **Decision Trees (DTree):** o árboles de decisión, donde cada nodo evalúa una característica y las hojas del árbol especifican la decisión a tomar. El algoritmo más utilizado y conocido se llama C4.5 [65], en su funcionamiento va (i) calculando la característica que proporciona la mayor ganancia de información en las muestras, (ii) crea un nodo de decisión utilizando el atributo que mejor divide el conjunto de datos de entrenamiento, (iii) crea listas secundarias de muestras utilizando los criterios de decisión creados y (iv) crea un Decision Trees (DTree) para todas las listas secundarias a partir del nodo de decisión. El algoritmo se detiene cuando todas las muestras en una lista secundaria pertenecen a una clase específica, que es cuando el algoritmo crea un nodo de decisión para esa clase. En [66] se encontró que los árboles de decisión podrían identificar a los sujetos con alta precisión usando datos del acelerómetro, pero utilizó una población muy pequeña de sólo 5 individuos.

Técnica ML	Referencia	Sensor	Modo	EER	H	FPR	Sujetos	Caract.
SVM	[Casale et al. 2012] [59]	ACC	V	-	-	0.01	20	18
	[Hestbek et al. 2012] [56]	ACC	I	0.1	-	-	36	12
	[Ho et al. 2012] [61]	ACC	V	-	1	-	36	-
	[Sugimori et al. 2011] [66]	ACC	I	-	0.98	-	5	2
GMM	[Lu et al. 2014] [34]	ACC	V	0.14	-	-	12	87
	[Meharia and Agrawal 2015] [33]	ACC	V	-	0.8	0.14	10	-
KNN	[Nickel et al. 2012] [67]	ACC	I	-	0.82	-	36	52
HMM	[Nickel et al. 2011] [68]	ACC	I	0.1	-	-	48	26
DTree	[Sugimori et al. 2011] [66]	ACC	I	-	0.98	-	5	2

Tabla 3.1: Resultados de artículos utilizando un smartphone.

En la tabla 3.1, las máquinas de vectores soporte consiguen buenos resultados, tanto en el

caso de identificación como de verificación, pero utilizando Bases de Datos muy pequeñas, con un máximo de 36 individuos. Destaca [66], ya que consigue una precisión alta de 0.98; y cuando se vuelve a utilizar con árboles de decisión consigue los mismos resultados, pero sigue ocurriendo lo mismo, con 5 individuos y 2 características no se pueden extraer muchas conclusiones. El modelo de Mixtura Gaussiana (GMM) parece obtener peores resultados que SVM, además llama la atención [34], ya que con 12 individuos se están usando 87 características, pudiendo existir un problema de sobreajuste que este perjudicando a los resultados. K -vecinos más próximos consigue una precisión de 0.82, peores resultados que SVM, pero de nuevo se están utilizando muchas más características, por lo que podría existir también aquí un problema de sobreajuste. Por último, en el modelo oculto de Markov (HMM) se obtiene una tasa de equierror razonable, de 0.1, y es la Base de Datos que contiene más sujetos, de un total de 48, con un número de características más pequeño que antes, de 26. Sin duda y como se puede ver en la tabla, el sensor más utilizado en la literatura es el acelerómetro, aunque hablan sobre el giroscopio, no lo utilizan tanto para obtener resultados. Pero existen artículos como [69], que comparan ambos sensores utilizando señales PPG durante la realización de ejercicio físico. Estas señales recogen el estado del corazón y otros órganos. Considerando que los acelerómetros por sí solos no pueden diferenciar entre aceleración debida al movimiento o a la gravedad y que las correlaciones presentes en las diferentes señales de movimiento (3-ejes del acelerómetro y los 3-ejes del giroscopio) recogen diferente información.

Un problema del uso de *smartphones* se encuentra en la ubicación del dispositivo, ya que no todos los usuarios lo llevan siempre en el mismo sitio y posición. Destaca el trabajo [48], donde se emplearon dispositivos portátiles diseñados específicamente para capturar los datos del movimiento mediante un acelerómetro, probando diferentes posiciones del dispositivo: en el pie, la cadera, el bolsillo del pantalón y la muñeca de los distintos usuarios, y se consiguieron los resultados de la tabla 3.2. Como en cada localización se utiliza distinto número de usuarios, se va a realizar una especie de EER por usuario, utilizando el cociente del EER y el número de individuos para poder compararlos. Parece que los mejores resultados se consiguen con el dispositivo en la cadera y en el bolsillo del pantalón, zonas muy similares, donde el acelerómetro podría capturar mejor el movimiento, mientras que en el tobillo los resultados son peores, probablemente porque sea una zona con más ruido, al ir demasiado en contacto con el pie y la muñeca parece capturar peor la información consiguiendo peores resultados. Pero hace falta recoger el EER medio utilizando las 4 localizaciones y el mismo número de individuos para que las conclusiones sean más realistas, ya que los 30 individuos usados para obtener el EER con el dispositivo en la muñeca pueden ser los que peores resultados estén dando por tener movimientos muy similares entre ellos.

No obstante, los sistemas biométricos rara vez alcanzan una precisión perfecta en la práctica debido a muchos factores, tales como el ruido, entrenamiento incompleto o un algoritmo de aprendizaje automática no ideal, lo cual afecta a la tasa obtenida de falsos positivos y falsos negativos. Todos estos resultados y la investigación realizada ha sido utilizando, la mayoría de las veces, smartphones o dispositivos creados para el propio propósito del estudio cuando en el presente proyecto se utilizan *wearables* comerciales, donde un algoritmo muy preciso podría drenar la batería rápidamente o tomar demasiado tiempo para tomar una decisión, lo cual no es factible.

Localización dispositivo	EER	Nº individuos	EER por individuo
Tobillo	5 %	21	0.24
Cadera	13 %	100	0.13
Bolsillo del pantalón	7.3 %	50	0.15
Muñeca	10 %	30	0.33

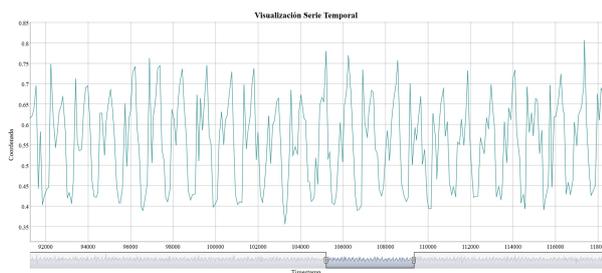
Tabla 3.2: Resultados cambiando la posición del dispositivo.

Capítulo 4

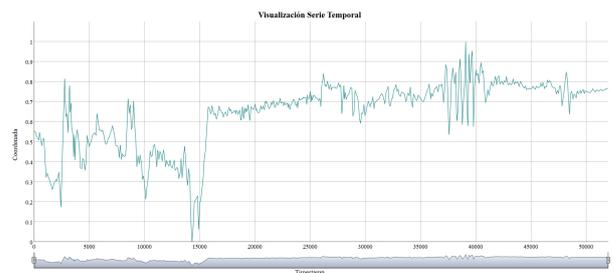
Datos y análisis

El conocimiento de los datos y su naturaleza requiere mucha cantidad de tiempo, pero son el primer paso para conseguir ubicarse correctamente en el problema que se quiere resolver. Para ello, en este capítulo se va a describir la base de datos que se va a utilizar a lo largo de todo el trabajo. También se va a realizar un análisis visual para encontrar cualquier anomalía existente en los datos, haciendo las correcciones de manera manual.

Respecto al análisis visual, idealmente se busca periodicidad, con un mismo patrón periódico que se repita a lo largo de las diferentes muestras del mismo usuario. Entre usuarios distintos, las series temporales deben ser también periódicas, pero con distinto patrón. Un ejemplo de serie periódica puede verse en la figura 4.1 (a). El caso opuesto sería una serie no periódica, cuyo ejemplo se puede ver en la figura 4.1 (b).



(a) Periódica



(b) No Periódica

Figura 4.1: Tipos posibles de series de datos.

4.1. Base de datos

Se ha utilizado una Base de Datos que había sido recogida durante la realización de un TFG anterior [4], con el objetivo de no perder tiempo en recoger nuevos datos y centrarse en el análisis de los ya disponibles. En ella se disponía de 21 usuarios. Cada uno realizó un recorrido andando

de minuto y medio, aproximadamente, durante 2 sesiones en diferentes días. Y dependiendo del usuario, cada día realizó el recorrido una única vez o dos.

En la tabla 4.1 se muestra la información disponible de cada uno de los usuarios: 13 hombres, 7 mujeres y 1 usuario sin identificar en edades comprendidas entre 16 y 57 años, utilizando el reloj y la pulsera en la mano dominante o la opuesta en función del usuario. En la misma tabla se muestra el número de datos recogidos de cada usuario. El número total de datos es 66.¹

*Usuarios que en lugar de realizar el recorrido una vez en cada sesión, lo realizaron las dos veces en la misma sesión: la primera.

Usuario	Sexo	Edad	Mano dominante	Mano portadora	Nº de datos
usuario0	-	-	-	-	2
usuario1	Hombre	21	Derecha	Izquierda	4
usuario2	Hombre	57	Derecha	Izquierda	4
usuario3	Hombre	50	Derecha	Izquierda	4
usuario4	Hombre	50	Derecha	Izquierda	2
usuario5	Mujer	53	Derecha	Izquierda	2*
usuario6	Hombre	21	Derecha	Derecha	2
usuario7	Mujer	16	Derecha	Izquierda	2
usuario8	Mujer	56	Derecha	Derecha	4
usuario9	Mujer	46	Derecha	Izquierda	4
usuario10	Mujer	19	Derecha	Izquierda	4
usuario11	Mujer	46	Derecha	Derecha	4
usuario12	Hombre	16	Derecha	Derecha	4
usuario13	Hombre	49	Derecha	Derecha	4
usuario14	Hombre	20	Derecha	Izquierda	4
usuario15	Hombre	22	Derecha	Derecha	4
usuario16	Mujer	48	Derecha	Izquierda	2
usuario17	Hombre	53	Derecha	Derecha	2*
usuario18	Hombre	22	Derecha	Izquierda	4
usuario19	Hombre	23	Derecha	Derecha	2
usuario20	Hombre	21	Derecha	Izquierda	2
TOTAL MUESTRAS DE DATOS DISPONIBLES					66

Tabla 4.1: Metadatos de los usuarios en la Base de Datos inicial.

Se dispone de la misma cantidad de datos en los 2 dispositivos comerciales y con 2 sensores en cada uno de ellos. Los dispositivos son un reloj *Motorola Moto 360* (Moto) y una pulsera *Microsoft Band 2* (Micro) que habían sido ya adquiridos para un TFG anterior [3] cuyo objetivo

¹El número de datos disponibles son 2 si realizó el recorrido una única vez en cada una de las dos sesiones o 4 si lo realizó dos veces, salvo las excepciones marcadas con asterisco.

fue desarrollar la aplicación móvil de recogida de datos, que fue empleada posteriormente en otro TFG [4], cuyos datos van a ser utilizados más ampliamente en este trabajo. Los dispositivos se pueden ver en la figura 4.2.



Figura 4.2: Dispositivos disponibles.

Los sensores utilizados son tanto el acelerómetro (ACC) como el giroscopio (GYR) tridimensional que poseen los dispositivos usados en la captura.

- **Acelerómetro:** mide la orientación de una plataforma fija respecto a la superficie terrestre. En esta situación podría verse como la rapidez con que algo se acelera.
- **Giroscopio:** mide la velocidad de rotación sobre un eje determinado.

Las 3 componentes son X, movimiento hacia la izquierda o derecha; Y, movimiento hacia adelante o hacia atrás; Z, movimiento hacia arriba o hacia abajo.

De manera resumida, al realizar cada recorrido, se va guardando en la Base de Datos la siguiente información.

- Identificador del usuario.
- International Mobile Equipment Identity (IMEI) del teléfono móvil o la herramienta utilizado para la adquisición de los datos. El IMEI es un código que identifica al aparato de forma exclusiva a nivel mundial.
- Dispositivo que se está utilizando (Micro o Moto).
- Tipo de sensor al que pertenece el dato (ACC o GYR).
- Timestamp: contiene tanto la fecha, como la hora con una precisión en milisegundos.
- Las coordenadas X, Y y Z del sensor indicado.
- Nombre del usuario.

- Número de la tarea, la sesión y la muestra para distinguir entre las diferentes tomas de datos del mismo usuario.

Con ello, se construye un fichero en formato CSV para cada toma de datos de cada usuario. El fichero contiene únicamente la información necesaria, que se va a utilizar a lo largo de este trabajo.

- Una primera columna con el tiempo relativo, que es la diferencia de tiempo entre una captura de las coordenadas X, Y, Z y la anterior. Los datos se almacenan con este valor temporal porque es más compacto que almacenar el timestamp.
- Tres columnas para las coordenadas X, Y, Z correspondientes a la captura de datos que marque el tiempo relativo. En la Base de datos tienen el nombre de dato1, dato2 y dato3 para hacer referencia a las coordenadas X, Y, Z respectivamente.

El recorrido dura, aproximadamente, minuto y medio, por lo que se tienen bastantes capturas para cada toma de datos de cada usuario. En la figura 4.3 se muestra un ejemplo de toma de datos, con el formato final con el que se va a trabajar.

	A	B	C	D	E
1	tiempoRelativo	dato1	dato2	dato3	
2	0	0,800781	-0,551514	0,602539	
3	68	1,062012	-0,692627	0,31665	
4	111	1,089844	-0,79248	0,296143	
5	102	0,954346	-0,553955	0,223877	
6	105	0,973145	-0,55835	0,19873	
7	94	1,097656	-0,663818	0,200439	
8	83	0,681885	-0,542725	0,268799	
9	148	1,078857	-0,369629	0,190674	
10	63	0,808838	-0,330078	0,204102	
11	71	0,764648	-0,465332	0,207764	
12	118	0,772461	-0,49585	0,30127	
13	99	0,766357	-0,543213	0,398438	
14	79	0,880615	-0,638672	0,471191	
15	136	0,775879	-0,593506	0,387939	
16	51	0,700195	-0,594727	0,356201	
17	101	0,719727	-0,477051	0,350586	
18	121	0,575928	-0,40625	0,341309	
19	79	0,584229	-0,411133	0,350098	

Figura 4.3: Formato de los datos que se van a utilizar.

4.2. Limpieza de los datos

La limpieza de los datos tiene el propósito de eliminar redundancias, inconsistencias, ruido o outliers, tratando de buscar las mejores decisiones para encontrar la mejor solución para el resultado final.

4.2.1. Pulsera Microsoft, acelerómetro (ACC):

Con el dispositivo Micro (Microsoft) y el acelerómetro, el usuario 13 presenta una señal mala para su segunda sesión y toma. Los gráficos correspondientes pueden verse en la figura 4.4 (coordenadas X, Y, Z de izquierda a derecha). La decisión ha sido eliminarla por tener una captura de datos errónea.

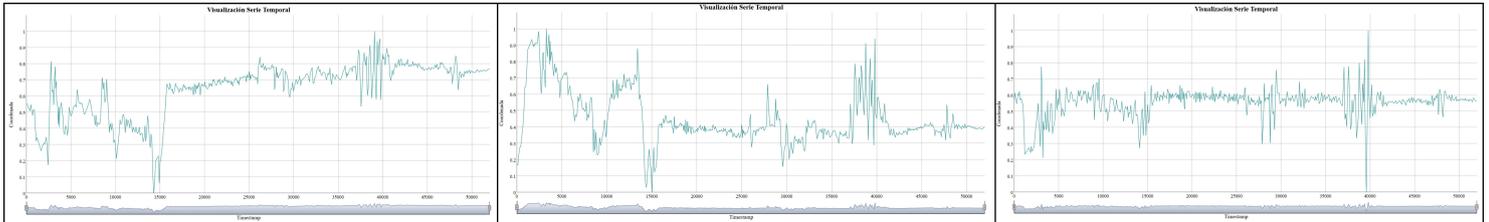


Figura 4.4: Usuario 13, sesión 2, toma 2, coordenadas X, Y, Z respectivamente.

Mientras en el resto de los usuarios no ocurría, el 19 y el 20 presentaban problemas en uno de sus dos registros de datos tomados. Como se puede ver en la figura 4.5, que corresponde con la serie de tiempo del usuario 19 en su componente Z, hay un problema en el “timestamp”: las muestras son almacenadas en el orden en que fueron tomadas, y, como se ve en la figura, hay zonas donde el timestamp no sigue este orden, produciéndose saltos temporales. No se sabe el porqué del problema, pero podría ser causado por el dispositivo o la herramienta de adquisición de los datos.

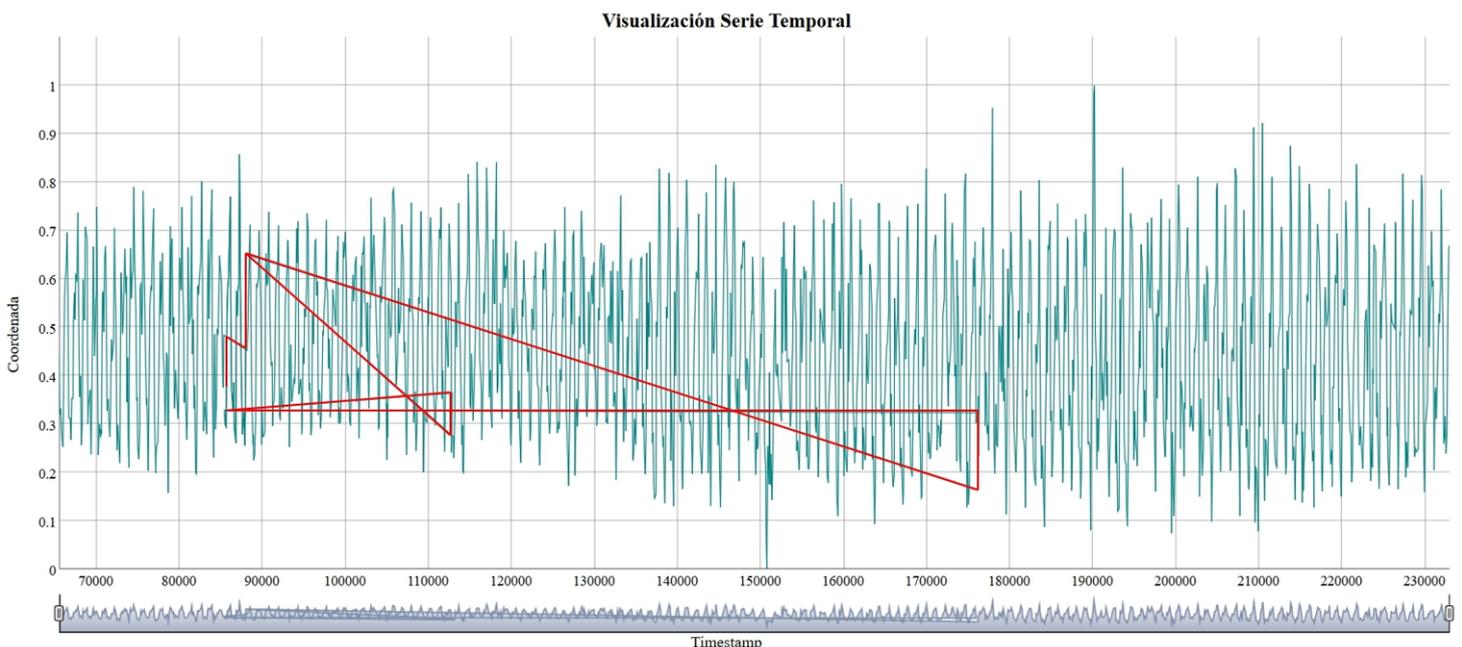


Figura 4.5: Mostrar error de retroceso en los datos.

La visualización de estos gráficos ha llevado a un análisis profundo de los datos numéricos originales. En ambos casos ocurre lo mismo, la presencia de valores negativos en el tiempo relativo,

es decir, al calcular la diferencia de tiempo entre un punto y el anterior. Además, mientras en todos los usuarios y registros, esta columna de valores empezaba en el 0, en el caso del usuario 19 empezaba en 65572 y en el usuario 20 en 33. En la figura 4.6 se muestra un ejemplo de valor negativo para el usuario 19.

	A	B	C	D	E	F
459	30	0,897949	-0,465576	0,126953		
460	78	1,058105	-0,434082	0,197754		
461	144	1,254395	-0,324951	0,126953		
462	87	1,223145	-0,484863	0,094971		
463	41	1,311035	-0,286133	-0,158447		
464	96	0,835449	-0,289307	-0,065674		
465	-27429	0,833008	-0,211914	-0,092529		
466	92	1,022705	-0,240479	-0,114746		
467	102	1,119141	-0,19751	-0,118164		

Figura 4.6: Error en el tiempo relativo del usuario 19.

Analizando detenidamente los datos, mediante una inspección visual de las zonas con problemas, vimos que si eliminábamos las muestras con tiempo relativo acumulado negativo, la señal se reconstruía perfectamente. Es decir, es como si se hubieran añadido puntos espurios, que bastaba con quitarlos para que la señal quedara correcta. Esta fue la solución que se adoptó, aplicando directamente en la hoja de cálculo el siguiente algoritmo:

1. Crear una columna con el valor de tiempo relativo acumulado.
2. Ordenar los datos de la tabla de menor a mayor usando la columna anterior creada.²
 - Si existen valores negativos, son los puntos espurios, se eliminan todas las filas que los contienen.
 - Si no existen valores negativos, no se elimina ninguna fila.
3. Eliminar la columna que contiene el tiempo relativo original.
4. Crear una nueva columna de tiempo relativo, pero ahora ya sin puntos erróneos. Empezando en valor 0 y restando los valores acumulados en ese momento con el anterior.
5. Eliminar la columna creada de acumulado, para dejar el fichero de datos con el mismo formato.

²De existir valores negativos tienen que estar al principio, ya que los datos están ordenados de menor a mayor.

En las figuras 4.7 y 4.9 se representan las 3 componentes del usuario 19 y 20 sin realizar ningún cambio. Mientras que tras aplicar el algoritmo anterior mencionado y representar de nuevo los datos, se obtienen los gráficos de las figuras 4.8 y 4.10. Comparándolos, se puede ver que la señal es la misma sin esos saltos. Se trata de señales, ambas periódicas, con un patrón más constante en el usuario 19.

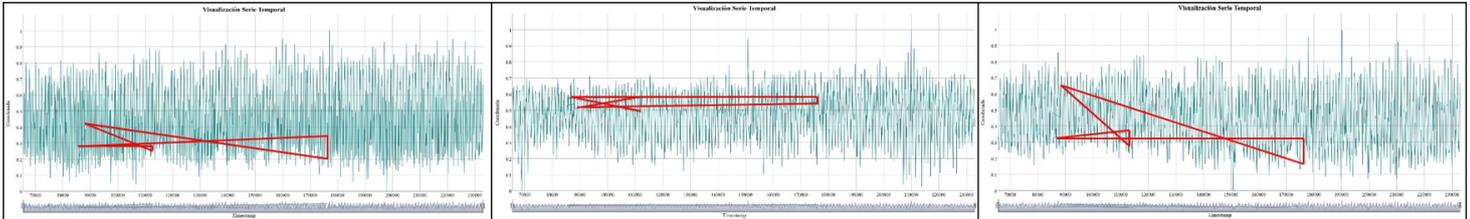


Figura 4.7: Usuario 19, sesión 2, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC)

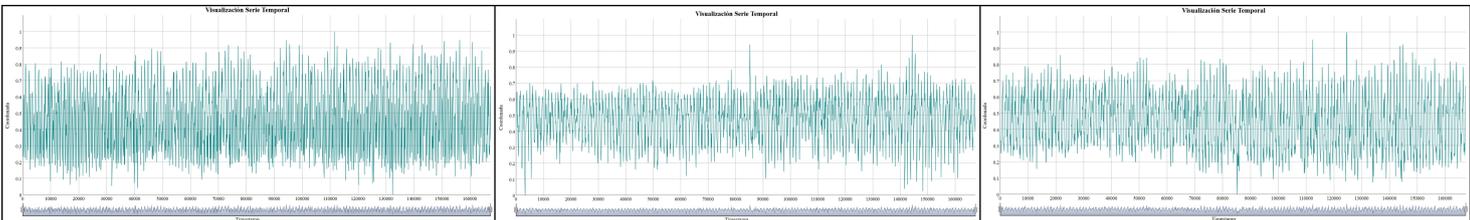


Figura 4.8: Usuario 19 modificado, sesión 2, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC).

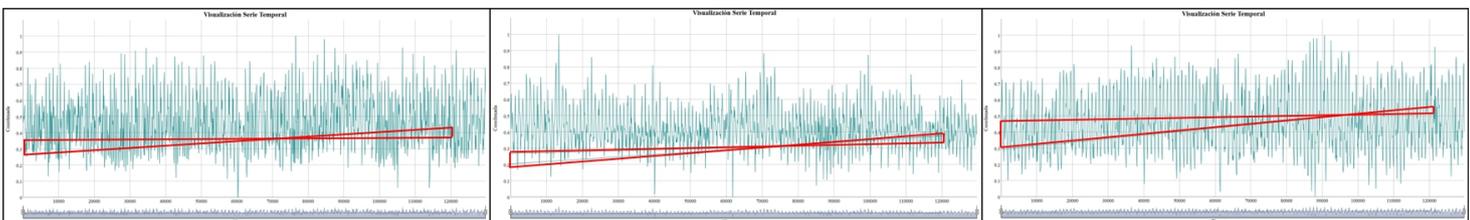


Figura 4.9: Usuario 20, sesión 1, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC)

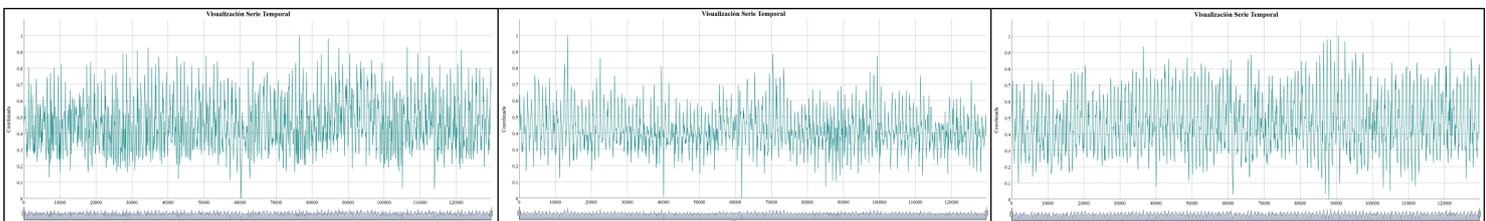


Figura 4.10: Usuario 20 modificado, sesión 1, toma 1, coordenadas X, Y, Z respectivamente (Micro ACC).

Se ha eliminado el usuario 0. Al principio no existían razones para hacerlo, salvo la pregunta de por qué en la tabla de los metadatos no existía información sobre dicho usuario. Los datos de partida han sido con labels ocultos, sin dar importancia al nombre del usuario, pero al visualizar

la carpeta de los datos originales (sin codificar la etiqueta) y codificados se observa como falta 1 dato, correspondiente a la sesión 1, muestra 1, del dispositivo Micro con el acelerómetro, que luego sí existe en los datos con la etiqueta codificada. Otra razón que ha motivado esta decisión es que al hacer el análisis visual y calcular la distancia euclídea entre las curvas que se construían con el Análisis de Fourier, de un usuario consigo mismo y con el resto, el usuario 0 era el único que presentaba una distancia *relativamente grande* consigo mismo como puede verse en la figura 4.11. Concluimos que fue un usuario ficticio de prueba de la aplicación móvil, por lo que los datos no se corresponden con una captura correcta.

X	us0_1	us0_2	us1_1	us1_2	us1_3	us1_4	us2_1	us2_2	us2_3	us2_4	us3_1	us3_2	us3_3	us3_4
us0_1	0	0,13	0,06	0,09	0,08	0,09	0,08	0,14	0,1	0,13	0,08	0,08	0,09	0,09
us0_2	0,13	0	0,07	0,08	0,01	0,11	0,11	0,17	0,13	0,16	0,03	0,02	0,03	0,11
us1_1	0,06	0,07	0	0,02	0,03	0,04	0,06	0,11	0,07	0,11	0,03	0,03	0,04	0,03
us1_2	0,09	0,08	0,02	0	0,05	0,04	0,08	0,13	0,09	0,13	0,03	0,03	0,04	0,09
us1_3	0,08	0,01	0,03	0,05	0	0,07	0,07	0,12	0,09	0,12	0	0,01	0,01	0,06
us1_4	0,09	0,11	0,04	0,04	0,07	0	0,07	0,13	0,09	0,12	0,05	0,04	0,07	0,09
us2_1	0,08	0,11	0,06	0,08	0,07	0,07	0	0,06	0,03	0,05	0,06	0,07	0,07	0,08
us2_2	0,14	0,17	0,11	0,13	0,12	0,13	0,06	0	0,02	0,02	0,12	0,12	0,13	0,14
us2_3	0,1	0,13	0,07	0,09	0,09	0,09	0,03	0,02	0	0,01	0,08	0,08	0,09	0,1
us2_4	0,13	0,16	0,11	0,13	0,12	0,12	0,05	0,02	0,01	0	0,11	0,12	0,12	0,13
us3_1	0,08	0,03	0,03	0,03	0	0,05	0,06	0,12	0,08	0,11	0	0	0	0,07
us3_2	0,08	0,02	0,03	0,03	0,01	0,04	0,07	0,12	0,08	0,12	0	0	0,01	0,08
us3_3	0,09	0,03	0,04	0,04	0,01	0,07	0,07	0,13	0,09	0,12	0	0,01	0	0,08
us3_4	0,09	0,11	0,03	0,09	0,06	0,09	0,08	0,14	0,1	0,13	0,07	0,08	0,08	0
us4_1	0,15	0,03	0,08	0,06	0,03	0,12	0,14	0,19	0,15	0,19	0,03	0,03	0,02	0,14

Figura 4.11: Distancias usuarios (componente X).

Los colores representan:

- Verde oscuro: valores de distancia entre 0 y 0.02.
- Verde claro: distancias entre 0.02 y 0.04.
- Naranja claro: distancias entre 0.04 y 0.06.
- Naranja oscuro: distancias entre 0.06 y 0.1.
- Rojo claro: distancias entre 0.1 y 0.15.
- Rojo oscuro: distancias entre 0.15 y 0.20.

Lo ideal con respecto a la figura 4.11 es que los cuadrados de la diagonal con borde negro tengan distancias pequeñas, es decir colores verde claro y oscuro o naranja claro, ya que están representando al mismo individuo y la distancia de un individuo con él mismo, idealmente tiene que ser pequeña. Y el resto de los elementos por fuera de la diagonal tengan distancias altas (colores naranjas oscuros, rojo claro y oscuro), ya que representan la distancia de un individuo con respecto al resto.

4.2.2. Pulsera Microsoft, giroscopio (GYR):

Al cambiar de sensor dentro del mismo dispositivo y conocido que en ciertas ocasiones, desconocidas podían ocurrir fallos en el orden de las muestras almacenadas apareciendo valores negativos en el tiempo relativo, era de esperar que volviese a ocurrir. Además, tratándose del mismo dispositivo, los fallos debían encontrarse en los mismos usuarios, ya que los datos de ambos sensores se recogen en el mismo instante de tiempo.

Efectivamente, como puede verse en las figuras 4.12 y 4.13, correspondientes a la componente Z de la segunda muestra del usuario 19 y a la primera del usuario 20, existe un retroceso de los datos. La manera de solucionarlo ha sido la misma que en el otro sensor, y los resultados son los mismos sin esos saltos. En este caso, en el usuario 19 se tiene una señal periódica buena, con mucho ruido al inicio y al final y en el usuario 20 una señal también periódica, con menos ruido.

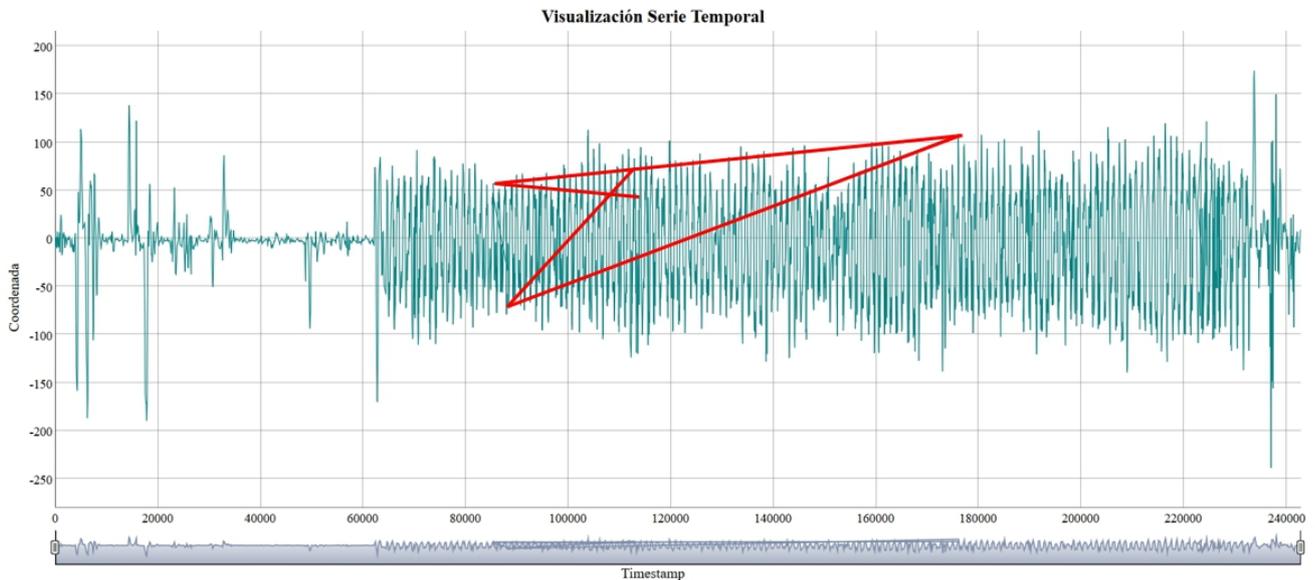


Figura 4.12: Usuario 19, sesión 2, toma 1, coordenadas Z (Micro GYR).

4.2.3. Reloj Motorola, acelerómetro (ACC):

Cambiando de dispositivo sigue ocurriendo el mismo problema de obtener valores negativos en algunos instantes de tiempo relativo que genera retrocesos y problemas en la señal de los datos. En este dispositivo ocurre en la primera sesión del usuario 4 y en la segunda del usuario 19, es decir, en el mismo número de muestras que antes, pero en un usuario diferente. Esto genera sospechas de que el problema se encuentra en la herramienta de adquisición de los datos y en que se está trabajando con dispositivos comerciales, donde las cosas no son tan ideales como cuando se trabaja con herramientas diseñadas para el propio propósito del trabajo.

El usuario 4 contiene al principio, un intervalo grande de tiempo donde no se han capturado datos y es al final, cuando de repente hay muchos valores negativos seguidos, lo que produce más

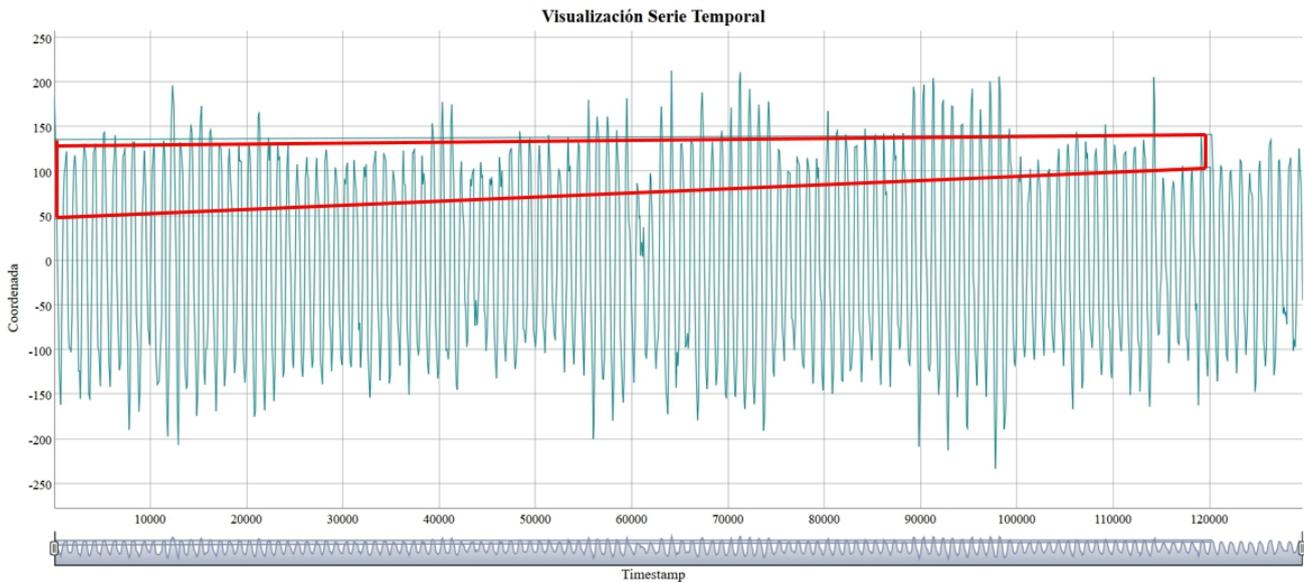


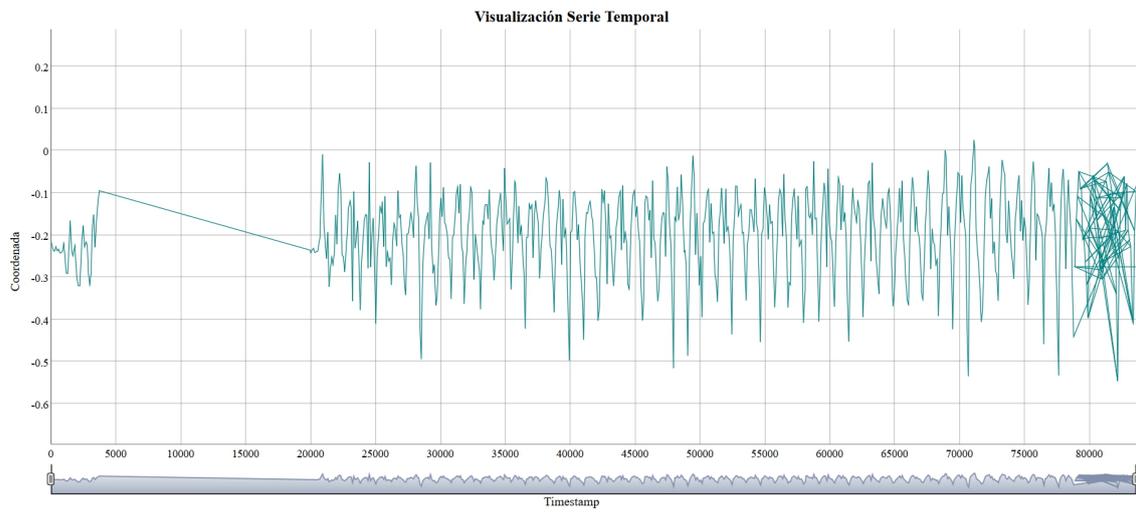
Figura 4.13: Usuario 20, sesión 1, toma 1, coordenadas Z (Micro GYR).

diferencias entre la señal errónea y corregida, como puede verse en la figura 4.14. En el usuario 19 hay menos valores negativos y su comportamiento es similar al de los usuarios en la pulsera de *Microsoft*.

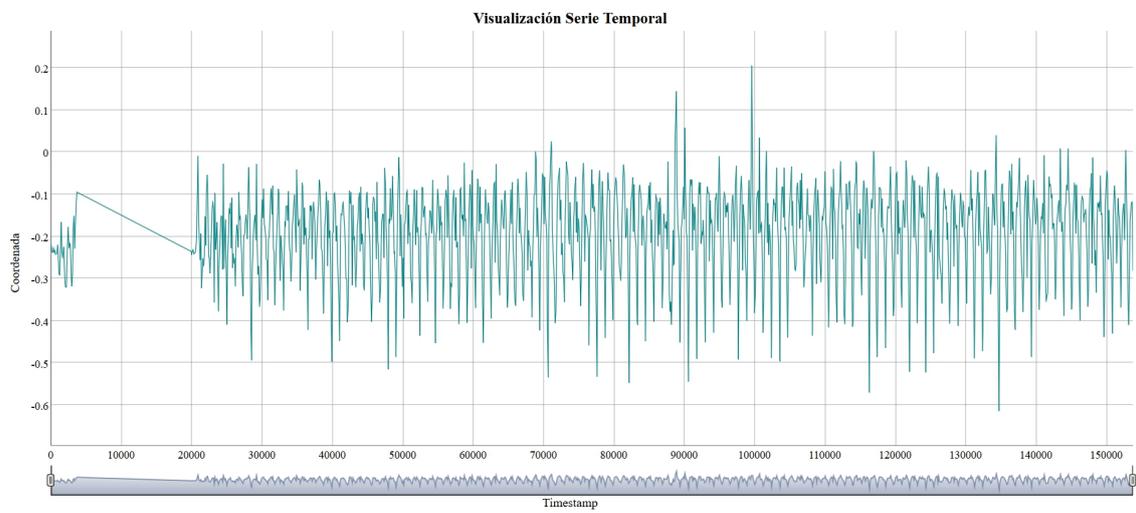
Por otro lado, a través de la visualización de la serie de los datos en cada usuario, se ha visto como en la última muestra del usuario 14, correspondiente a la segunda toma de la sesión 2, se tiene una señal periódica, bastante buena, pero con una duración mucho más pequeña a la del resto de usuarios, e incluso a la del mismo usuario en las otras tomas. Además, con una diferencia bastante grande, ya que en la primera toma de la sesión 2 tarda aproximadamente 120.000 milisegundos en realizar el recorrido; cuando en la segunda toma tarda 50.000 milisegundos; y en la primera toma de la primera sesión 170.000, 3 veces más. Esto podría producir problemas en las decisiones que se tomen para construir el sistema de reconocimiento final, si se decide trabajar con tamaños grandes.

4.2.4. Reloj Motorola, giroscopio (GYR):

En el otro sensor del reloj Motorola ocurre lo mismo, ya que se trata de usuarios andando en el mismo instante de tiempo con el reloj situado en la muñeca, mientras una herramienta está adquiriendo los datos que están generando los sensores acelerómetro y giroscopio, por lo que los problemas surgidos son los mismos. Los usuarios 4 y 19 presentan valores negativos de tiempo relativo siguiendo el mismo comportamiento que en el acelerómetro del mismo dispositivo, reloj *Motorola*. De la misma manera, el usuario 14 sigue presentando una toma de datos buena pero muy pequeña en la segunda muestra de la sesión 2.



(a) Señal errónea



(b) Señal corregida

Figura 4.14: Señal usuario 4, componente Y, Motorola Acelerómetro.

Capítulo 5

Configuración experimental

Tras conocer más a fondo los datos que se van a utilizar y realizar una limpieza manual de los mismos en el capítulo anterior, aquí se van a empezar a tomar decisiones de los parámetros estáticos existentes en el problema, que se utilizarán en los posteriores capítulos.

Como este trabajo se ha realizado de manera conjunta con otro del Grado en Estadística [5], aquí no se incluye pero se ha realizado un análisis de la autocorrelación de los datos con el objetivo de conseguir un análisis más profundo de los mismos y comprobar si existía periodicidad o no en la señal. En ese trabajo, se trataba el cálculo de la autocorrelación como si se dividiese la señal en ventanas de tamaño k muestras y se obtuviera cuánto se parecían esas ventanas entre sí. Para ello, ya era conocido a priori que la duración de un paso andando era de aproximadamente un segundo y se demostraba consiguiendo los valores más altos de autocorrelación en tamaños de ventana, en milisegundos, entre 830 y 1245.

El tamaño de la ventana es un parámetro importante que se irá analizando en los siguientes capítulos. Representa el tamaño de la señal usada para reconocer al usuario. Como de cada individuo hemos adquirido distintas muestras de duración aproximada un minuto y medio, había dos maneras de abordar el problema: usando toda la muestra adquirida o dividiendo esa muestra en marcos temporales. Aquí se va a trabajar con la extracción de características, por lo que, en el primer caso, se obtendría un único vector de características por muestra, mientras que en el segundo, una secuencia de ellos. Se ha decidido ir por la segunda opción para capturar la naturaleza variable del rasgo biométrico.

En definitiva, con el estudio del análisis de la autocorrelación realizado a la señal completa, sin eliminar el ruido ni realizar ningún tipo de preprocesamiento se han conseguido las siguientes conclusiones:

- A mayor tamaño de ventana, en general, o se mantienen o se obtienen mejores resultados, consiguiéndose autocorrelaciones más altas y por tanto ventanas de usuarios más periódicas.
- Existe un límite en el tamaño de la ventana a partir del cual ya no se observan mejoras en los valores de autocorrelación, pudiendo empeorar en algunas componentes.

- Una combinación de los dos sensores utilizados, acelerómetro y giroscopio podría ser apropiada. Ya que en los dos dispositivos se ha observado lo mismo, mejores resultados en la componente Y y Z del giroscopio y en la componente X y el módulo del acelerómetro.

Relacionando una señal periódica con una señal con coeficiente de autocorrelación en valor absoluto alto y, en consecuencia, una señal que tiene un patrón y que es susceptible de poder ser usada en biometría.

La configuración experimental está muy relacionada con los objetivos del trabajo. No se pretende, con este proyecto, llegar a un sistema de reconocimiento biométrico basado en ponibles, si no, conseguir una primera aproximación al problema en la que se analicen los distintos elementos que entran en juego en el sistema, para entenderlos mejor y ver su relación en el rendimiento del reconocimiento del usuario. En definitiva, se busca plantar unas bases sólidas sobre las que seguir trabajando, ya sí, en un sistema eficiente de reconocimiento.

5.1. Extracción de características

Tras un análisis detallado de los artículos y la documentación actual más relevante en el campo de la biometría, y en concreto de los dispositivos ponibles, resumen de ello se puede encontrar en el capítulo 3 del Estado del Arte, se ha decidido dividir la señal en marcos temporales o ventanas, y sobre cada una de ellas, extraer características en el dominio del tiempo y en el dominio de la frecuencia, para estudiar las diferencias. Hasta donde conocemos, nunca se ha realizado este estudio comparativo.

Posteriormente, cuando se estudie el rendimiento final, se estudiará la influencia de hacer una selección de características para eliminar aquellas que sean irrelevantes, redundantes o altamente correlacionadas entre sí con el objetivo de mejorar los resultados y eliminar posible sobreajuste.

5.1.1. Dominio del tiempo

Los datos capturados son muestras de la evolución de una señal en el tiempo, donde se ha obtenido para cada toma de datos de cada usuario el instante de tiempo (*timestamp*) y los datos en los sensores. Tras un análisis bibliográfico de las características extraídas directamente sobre estas señales, es decir, en el dominio del tiempo se decidió probar las siguientes:

- Periodo de cada una de las componentes en el que se consigue la máxima autocorrelación entre las pruebas hechas mencionadas en el apartado del análisis de la autocorrelación del trabajo [5].
- Valor de la autocorrelación en cada una de las componentes donde se consigue el máximo.
- Medidas estadísticas en cada una de las componentes (X/Y/Z):

- Media: medida de tendencia central que representa el centro de gravedad de la distribución de la variable.
 - Mediana: medida de tendencia central que representa el valor de la variable en la posición central entre un conjunto de datos ordenados.
 - Máximo: el valor más grande entre el conjunto de valores.
 - Mínimo: el valor más pequeño entre el conjunto de valores.
 - Desviación estándar: medida de dispersión que se utiliza para cuantificar la variación de un conjunto de datos.
 - Rango: intervalo entre el valor máximo y el valor mínimo, proporcionando una idea de la dispersión de los datos.
 - Kurtosis: característica de forma de la distribución de probabilidad/frecuencias de los datos que indica que tan apuntada o achatada se encuentra una distribución respecto a un comportamiento normal (distribución normal). Valores grandes indican mayor concentración de valores de la variable tanto muy cerca de la media de la distribución (pico) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios, no implicando con ello una mayor varianza [70].
 - Quantil 25 % y 75 %: puntos tomados a intervalos regulares de la función de distribución de la variable aleatoria. Lo que se ha utilizado es dividir la distribución en cuatro partes correspondientes a los cuantiles 25 %, 50 % (media) y 75 %.
 - Coefficiente de asimetría: representa el grado de simetría (o asimetría) de la distribución de probabilidad de la variable aleatoria. Considerando como eje de simetría la recta paralela al eje de ordenadas que pasa por la media de la distribución, una distribución es simétrica si existe el mismo número de valores a la derecha que a la izquierda de la media y por tanto el mismo número de desviaciones con signo positivo que con signo negativo. Mientras que hay asimetría positiva si hay valores más separados de la media por la derecha y asimetría negativa si hay valores más separados de la media por la izquierda [71].
- Energía conjunta de las 3 componentes: proporciona una idea de la disposición de los datos tridimensional.

$$Energia = \frac{1}{N} \cdot \sum_{n=1}^N (\sqrt{X[n]^2 + Y[n]^2 + Z[n]^2})^2 \quad (5.1)$$

- Ratio medio de las componentes XY: media de todos los cocientes de X e Y. Proporciona una medida de la gravedad de la distribución conjunta de ambas componentes de la variable.
- Ratio medio de las componentes XZ: media de todos los cocientes de X y Z.
- Ratio medio de las componentes YZ: media de todos los cocientes de Y y Z.

5.1.2. Dominio de la frecuencia

Para pasar al dominio de la frecuencia se ha usado, como es habitual, la transformada de Fourier.

La transformada de Fourier (FT) descompone la función del tiempo original en las frecuencias que la constituyen. Es un conjunto de números complejos, cuya magnitud (módulo) representa la amplitud de cada frecuencia presente en la función original y cuyo argumento es el desfase de la onda.

Para que los resultados de la transformada de Fourier sean interpretables, es decir, se pueda calcular la frecuencia de cada componente extraído, la señal original debe estar muestreada a frecuencia constante. De estudios realizados en trabajos anteriores [4], se vio que muestrear a más de 12 Hz. era innecesario; se fijó ese valor para la frecuencia de muestreo. El problema que tenemos con la señal capturada es que, al ser dispositivos reales, el periodo entre dos muestras consecutivas no es constante. Por lo tanto, lo primero que se tuvo que hacer para aplicar la FT es un remuestreo de la señal a 12 Hz. La bibliografía muestra distintas alternativas, de las cuales la más sencilla es usar la interpolación lineal, que se explicará en el capítulo de preprocesamiento de [5]. Alternativa que suele dar un rendimiento similar o superior a algoritmos más complejos, por lo que fue la técnica de remuestreo usada aquí.

La ecuación (5.2) muestra la expresión de la Transformada de Fourier Discreta (TDF) [72].

$$X_k = \sum_{n=0}^{N-1} X_n \exp\left(\frac{-i2\pi kn}{N}\right) \quad (5.2)$$

Donde:

- X_k : cantidad de frecuencia k en la señal; cada valor k^{th} es un número complejo que incluye amplitud (fuerza) y cambio de fase.
- N : número de muestras.
- n : muestra, $n \in \{0 \dots N - 1\}$.
- k : frecuencia entre 0 Hz. y $N-1$ Hz.
- $1/N$: tamaño real de los picos de tiempo.
- n/N : porcentaje de tiempo que ha pasado.
- $2\pi k$: velocidad en *radianes/segundo*.
- \exp^{-ix} : camino circular hacia atrás que indica cuánto nos hemos movido, para esta velocidad y tiempo.

El resultado de la TDF, X_k , es un número complejo del que nos interesa su amplitud (la amplitud carece de información biométrica) y la frecuencia asociada. Ambos valores se calculan de la siguiente manera:

- *Amplitud* = $\sqrt{\text{Re}(X_k)^2 + \text{Im}(X_k)^2}$ donde Re e Im son la parte real e imaginaria del número complejo X_k .
- *Frecuencia* = $0 : (\text{length}(X_k) - 1) * f / \text{length}(X_k)$

La teoría dice que la división de una señal en ventanas tiene un efecto sobre sus componentes frecuenciales. Este efecto es mayor cuando la TDF se aplica sobre la ventana extraída directamente de la señal. Para suavizar o paliar este efecto, se suelen usar las denominadas *funciones ventana*. Estas son funciones matemáticas que tratan de evitar discontinuidades al principio y al final de la señal. Al multiplicar la señal $s(t)$ por una función ventana $h(t)$, se genera $S_h(t) = s(t) \cdot h(t)$ que será la señal sobre la que se aplique la TDF.

Hay muchos tipos de ventanas, las más usadas son:

- Hamming: se define con la función matemática de (5.3) y tiene la forma que muestra la figura 5.1 (a).

$$v(n) = 0.53836 - 0.46164 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \quad (5.3)$$

- Hanning: se define con la función matemática de (5.4) y tiene la forma que muestra la figura 5.1 (b).

$$v(n) = 0.5 - 0.5 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \quad (5.4)$$

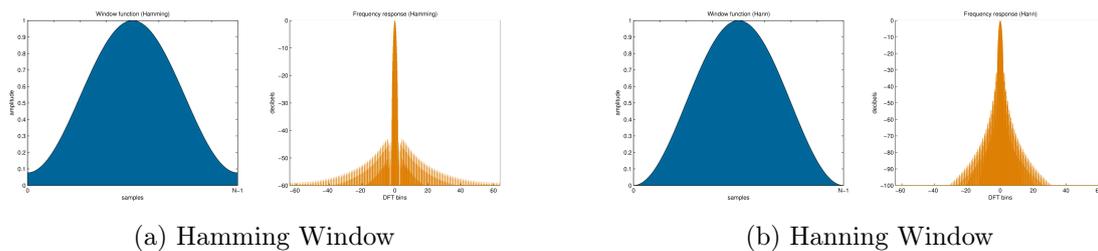


Figura 5.1: Ventanas de suavizado aplicadas a la transformada de Fourier.

Tras analizar las gráficas de Fourier sin aplicar las ventanas de suavizado y después aplicándolas, se ha podido ver como no existían diferencias significativas. Un ejemplo de ello puede verse en la figura 5.2. Por esta razón, la decisión ha sido no usar estas ventanas.

Para eliminar la componente de continua (componente 0 de la TDF), que no aporta información, se ha aplicado el filtro de restar la media en cada ventana, tal y como hace el artículo [44]. A partir de lo indicado en la bibliografía, se han extraído las siguientes características en cada una de las componentes X/Y/Z:

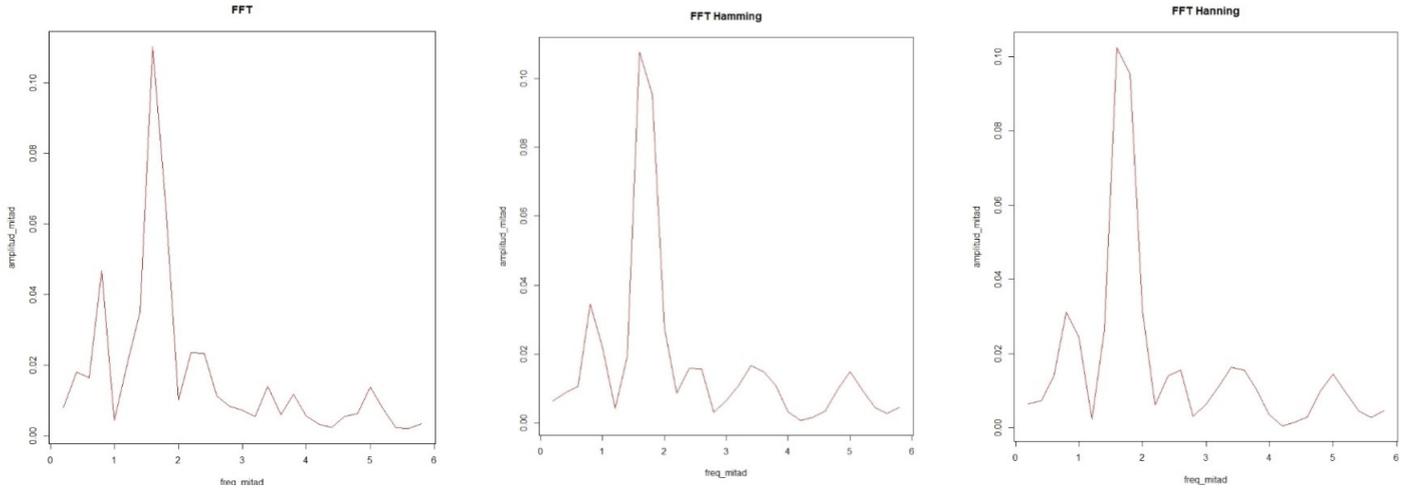


Figura 5.2: Aplicación de Fourier con y sin ventanas al usuario1, S1, M1, 1^a ventana

- Primera y segunda amplitud dominante: representa los dos valores más altos obtenidos entre las amplitudes resultantes del Análisis de la transformada de Fourier en cada una de las componentes de los datos.
- Primera y segunda frecuencia dominante: representa los dos valores de la frecuencia correspondientes a los dos puntos donde se consiguen las amplitudes anteriores.
- Área bajo la curva de Fourier (AUC) basado en splines: utiliza una interpolación de splines para calcular la cantidad de área bajo la curva formada por las amplitudes del Análisis de Fourier.
- Las mismas medidas estadísticas que en el dominio del tiempo, quitando el máximo y el mínimo.

5.1.3. Señal combinada

Se ha trabajado combinando la señal a través del módulo (5.5), tal y como se hace en la bibliografía [47, 48, 50, 56]. Una vez aplicado el módulo, se extraerán las características mostradas tanto en el dominio del tiempo como de la frecuencia.

$$\text{Modulo} = \sqrt{X^2 + Y^2 + Z^2} \quad (5.5)$$

Otra alternativa, menos utilizada, es el uso del arcoseno (5.6) [52, 55]. La bibliografía muestra resultados similares al módulo, por lo que fue la alternativa probada en el presente trabajo.

$$\text{Arcoseno} = \arcsin \frac{Z}{\sqrt{X^2 + Y^2 + Z^2}} = \arcsin \frac{Z}{\text{Modulo}} \quad (5.6)$$

Otra alternativa que se barajó y estudio en este proyecto fue fusionar las coordenadas a nivel de características, es decir, creando un vector de características resultante de juntar las de las coordenadas X, Y y Z. Esto nos daba un vector de 79 características. Las pruebas prospectivas realizadas no mostraron un buen rendimiento de esta alternativa, que, junto con el más alto coste computacional debido al mayor tamaño del vector de características, nos hizo desechar esta vía de trabajo.

5.2. Medición del error

Otra decisión importante es cómo evaluar los modelos implementados con el objetivo de poder compararlos y buscar la mejor solución final.

Entre las medidas más utilizadas en los sistemas biométricos se encuentran las curvas ROC (*Receiver Operating Characteristic*), éstas son una representación gráfica de la sensibilidad frente a la especificidad para un sistema de clasificación binario según se varía el umbral de decisión.

Nuestro problema se corresponde con el de clasificación binaria, dado que, para cada usuario, se considera a dicho usuario como auténtico y al resto como usuarios impostores.

Las medidas de error básicas usadas en este tipo de problemas son:

- Falsos positivos (False Positives o FP) o falsa aceptación: ocurre cuando se identifica a una persona no autorizada como autorizada. De manera que, si el sistema trata de verificar la identidad de una persona, un usuario impostor podría acceder de forma no autorizada.
- Falsos negativos (False Negatives o FN) o falso rechazo: ocurre cuando se impide el acceso a una persona autorizada.
- Verdaderos positivos (True Positives o TP): ocurre cuando el sistema trata de verificar la identidad de una persona y un usuario auténtico (verdadero) accede de forma correcta y es autorizada.
- Negativos verdaderos (True Negatives o TN): ocurre cuando el sistema trata de verificar la identidad de una persona y un usuario impostor es rechazado.
- Sensibilidad (True Positive Rate o TPR): proporción de usuarios auténticos que se consideran correctamente como autorizados, con respecto a todos los usuarios auténticos. En función de los términos anteriores, se puede calcular con la fórmula (5.7).

$$\text{Sensibilidad} = \frac{\text{TruePositive}}{\text{FalseNegative} + \text{TruePositive}} = \frac{TP}{FN + TP} \quad (5.7)$$

- Especificidad (False Positive Rate o FPR): proporción de usuarios impostores que se consideran erróneamente como autorizados con respecto a todos los usuarios impostores, cuyo resultado se puede obtener con la fórmula (5.8).

$$\text{Especificidad} = \frac{\text{FalsePositive}}{\text{FalsePositive} + \text{TrueNegative}} = \frac{FP}{FP + TN} \quad (5.8)$$

Tanto la sensibilidad como la especificidad tienen valores en el rango $[0,1]$, generando una curva ROC en estos rangos donde su área se denomina AUC. Los valores de AUC se interpretan de manera que cuanto mayor sea el valor del AUC, mejor es el rendimiento del modelo.

Otra medida del rendimiento muy utilizada en biometría es la tasa de equierror, que es el punto de intersección entre ambas tasas: sensibilidad y especificidad, conocido como *Equal Error Rate (EER)*. Cuanto menor sea su valor, mejor será el sistema.

La figura 5.3 muestra la especificidad en el eje de abscisas y la sensibilidad en el eje de ordenadas, generando la curva sobre su área (AUC) marcado en gris. El valor de la tasa de equierror se produce con $FPR=0.2$ y $TPR=0.8$.

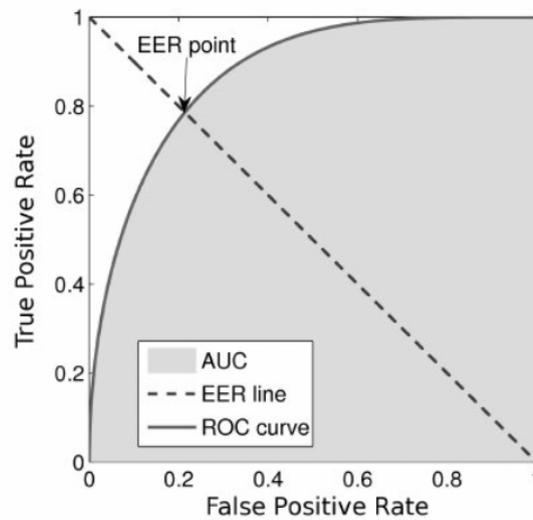


Figura 5.3: Ejemplo de EER a partir de la curva ROC y el AUC.

Como resultado final, tenemos dos maneras de mostrar el error:

- De manera gráfica: como se muestra en la figura 5.3, los valores de la sensibilidad y la especificidad para distintos valores umbrales.
- Mediante valor numérico: utilizando el área bajo la curva ROC o la tasa de equierror explicada. Pero estos valores se pueden calcular de manera individual para cada usuario o de manera global como la media de todos los usuarios disponibles.

Dado que nuestro objetivo aquí es comparar resultados, la opción gráfica es poco práctica en este caso, siendo la más habitualmente utilizada, y la que decidimos adoptar, la de utilizar valores numéricos, considerando ambas métricas, aunque mayoritariamente se va a utilizar el EER. Para la toma de decisiones se utilizará su valor medio con respecto a todos los usuarios, pero posteriormente se aprovechará la ventaja de tener pocos usuarios para hacer un estudio detallado de cada uno de ellos.

5.3. Experimentos

Teniendo en cuenta el contenido de la base de datos o corpus que estamos usando, se tienen:

- Diversos **usuarios**.
- Dos **sesiones** posibles en que se recogieron datos. S1 y S2 hacen referencia a la sesión 1 y 2 respectivamente.
- Un máximo de dos **muestras** de datos tomadas por sesión y pulsera a cada usuario, representándose como M1 y M2 para referenciar a la muestra 1 y 2 respectivamente.

Para cada usuario i vamos a tener los siguientes conjuntos de datos:

- Conjunto de entrenamiento (*train*): contiene los datos del usuario auténtico que se usarán para crear su patrón.
- Conjunto de prueba (*test*): distinguiendo entre:
 - **Muestras auténticas**: Serán muestras del usuario distintas a las usadas para el entrenamiento. Se usarán para calcular la tasa de falsos negativos.
 - **Muestras impostores**: Serán muestras de otros usuarios distintos al usuario i . Simularán ataques al sistema, por lo tanto, se usarán para calcular la tasa de falsos positivos.

Con respecto a la sesión y muestra, tenemos las siguientes pruebas:

1. Monosesión-Monomuestra (MonoMono): compara los datos dentro de la misma sesión y muestra, es decir, las muestras usadas para entrenamiento y para *prueba auténtico* del usuario son tomadas en la misma sesión. Es el caso más favorable y el que primero abordaremos para analizar los parámetros del sistema.
 - **Train**: S1, M1, usuario i
 - **Test**:
 - Test Auténticos: S1, M2, usuario i
 - Test Impostores: S1, M2, usuario $j \neq i$
2. Multisesión-Monomuestra (MultiMono): Las muestras usadas para entrenamiento y *prueba auténtica* son tomadas en distintas sesiones. Aquí se quiere probar la variabilidad del rasgo biométrico con el tiempo.
 - **Train**: S1, M1, usuario i
 - **Test**:

- Test Auténticos: S2, M1, usuario i
 - Test Impostores: S1, M2, usuario $j \neq i$
3. Multisesión-Multimuestra (MultiMulti): En biometría se ha demostrado que la variabilidad del rasgo con el tiempo es un problema que afecta al rendimiento del sistema. Una forma de paliarlo es intentar incluir en el modelo del usuario esta variabilidad. Una manera de hacerlo es usar para entrenamiento muestras de distintas sesiones. Esto es lo que se prueba aquí.
- **Train**: S1 y S2, M1, usuario i
 - **Test**:
 - Test Auténticos: S1 y S2, M2, usuario i
 - Test Impostores: S1 y S2, M2, usuario $j \neq i$

El protocolo experimental seguido para cada caso, *Monosesión-Monomuestra*, *Multisesión-Monomuestra* y *Multisesión-Multimuestra*, es el indicado. Ahora bien, otra forma que puede parecer más razonable de actuar es considerar las distintas posibilidades dentro de *Monosesión-Monomuestra*, que serían la indicada junto con las siguientes tres:

1. **Train**: S2, M1, usuario i
 - Test Auténticos: S2, M2, usuario i
 - Test Impostores: S2, M2, usuario $j \neq i$
2. **Train**: S1, M2, usuario i
 - Test Auténticos: S1, M1, usuario i
 - Test Impostores: S1, M1, usuario $j \neq i$
3. **Train**: S2, M2, usuario i
 - Test Auténticos: S2, M1, usuario i
 - Test Impostores: S2, M1, usuario $j \neq i$

De la misma forma, en *Multisesión-Monomuestra* existirían además de la mencionada las siguientes tres:

1. **Train**: S1, M2, usuario i
 - Test Auténticos: S2, M2, usuario i
 - Test Impostores: S1, M1, usuario $j \neq i$
2. **Train**: S2, M1, usuario i
 - Test Auténticos: S1, M1, usuario i

- Test Impostores: S2, M2, usuario $j \neq i$

3. **Train**: S2, M2, usuario i

- Test Auténticos: S1, M2, usuario i
- Test Impostores: S2, M1, usuario $j \neq i$

Y por último, en *Multisesión-Multimuestra* existiría, además de la mencionada otra más que es:

1. **Train**: S1 y S2, M2, usuario i

- Test Auténticos: S1 y S2, M1, usuario i
- Test Impostores: S1 y S2, M1, usuario $j \neq i$

Utilizando todas las posibilidades se haría una especie de validación cruzada. Esto, que parece una buena idea, no lo pudimos hacer debido a las deficiencias de los datos de que disponemos, ya que no todos los usuarios tienen todas las sesiones y dos muestras en cada sesión. Actuar de esa manera nos obligaría a quedarnos solo con los que tienen todo, lo que supone un subconjunto demasiado pequeño. En nuestros datos se tienen: 11 usuarios completos, 2 usuarios con solo una sesión con dos muestras, 7 usuarios con 1 sola muestra en cada sesión y 1 usuario con 3 datos, dos de la primera sesión y sólo uno de la segunda.

Por otro lado, entre todas las posibilidades, se ha selecciona la primera opción del procedimiento experimental mencionado, que utiliza como entrenamiento la primera sesión y muestra, siguiendo las pautas y bases fijadas en biometría que intentan simular el comportamiento de la vida real. Interpretando que los datos se utilizan en orden y que la primera sesión S1 y muestra M1 que se obtiene es la que forma parte del conjunto de *train*, la que se usa para lo que en biometría se denomina *inscribir* al usuario.

Los sensores utilizados son tanto el acelerómetro como el giroscopio tridimensional que poseen los dispositivos usados en la captura. Esto permite 4 posibilidades (*Microsoft acelerómetro*, *Microsoft giroscopio*, *Motorola acelerómetro*, *Motorola giroscopio*), de las cuales se pueden comparar si los resultados de los sensores son similares y apropiados para trabajar de manera complementaria o si sería mejor centrarse en uno; igual que en los dispositivos, para poder ver si las conclusiones se pueden generalizar y existen posibilidades de encontrar un sistema de reconocimiento bueno para cualquier dispositivo comercial.

5.4. Clasificación

Cuando se tiene un problema, se tiende a utilizar muchos algoritmos de clasificación diferentes o con pequeñas variaciones para conseguir resolverlo obteniendo el mejor resultado final. En este

trabajo, no se pretende resolver el problema completo, sino realizar un buen estudio prospectivo que sienta las bases de esta novedosa biometría, y dado que existen muchas incógnitas: *¿con qué tamaño de ventana extraer características?*, *¿qué características extraer?*, *¿qué tipo de preprocesado beneficiará más a los datos?*, *¿qué valor umbral fijar para permitir ventanas con autocorrelación alta?*, se ha elegido un clasificador sencillo que no introduce muchos parámetros al problema.

La elección es el algoritmo basado en distancias de **k-vecinos más próximos**. Un método simple, fácil de programar y entender si se necesita explicar a un público amplio. Además, solo necesita muestras del usuario para crear su patrón; la mayoría de los clasificadores discriminantes necesitan para su entrenamiento muestras de la clase auténtica y de la clase impostor, lo que introduce la variabilidad asociado a qué muestras de la clase impostor usar. Para nuestro problema, se puede resumir su funcionamiento a través de un bucle de usuarios auténticos y otro de usuarios impostores.

Partiendo de los datos divididos en *conjunto de entrenamiento* y *conjunto de prueba*, tal como se explica en la subsección 5.3 y con el objetivo de encontrar las distancias correspondientes a los usuarios auténticos, se tendrán que seguir los siguientes pasos *para cada usuario i* a estudiar:

1. Para cada ventana en el conjunto de datos de prueba del usuario i , calcular la distancia entre esta ventana y cada una de las ventanas en el conjunto de entrenamiento del usuario i que se esté estudiando.
2. Con las distancias obtenidas en el paso anterior, seleccionar las k distancias más pequeñas. Como distancia se ha usado la euclídea. Se pueden utilizar otras distancias, pero está es la más general y típicamente usada.
3. La distancia final o “score” de la muestra de prueba se obtiene mediante un estadístico (media, mediana, etc.) sobre las k distancias del paso 2.

Con el bucle anterior, se obtienen los scores para el conjunto de muestras auténticos, pero hay que realizar otro bucle que obtenga las distancias de los usuarios impostores j . En este bucle, se repiten las mismas operaciones que en el anterior, pero ahora, para las muestras de prueba del resto de usuarios, es decir, $\forall j \neq i$.

El pseudocódigo de ambos bucles puede verse en los algoritmos 1 y 2, donde las tablas de entrada del conjunto de datos *train* y *test* contienen los atributos de interés, eliminando aquellas columnas que indican el usuario, la sesión y la muestra, ya que se va calculando una distancia entre variables, todas ellas, numéricas y se está trabajando con datos de la misma clase, sin indicar la variable respuesta.

La función *distanciaEuclidea*(x, y) es la que calcula la distancia euclídea entre dos vectores con n características numéricas siguiendo (5.9).

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (5.9)$$

Algorithm 1: Algoritmo para obtener los scores de los usuarios auténticos

Input Tabla de datos de entrenamiento de todos los usuarios $train_i$
 Tabla de datos de prueba de todos los usuarios $test_i$
 Usuarios disponibles en la forma experimental $usuariosDisponibles$
Output Scores auténticos de todos los usuarios $scoreAutentico_i$

```

for  $user$  in  $usuariosDisponibles$  do
   $n\_ventanas\_train \leftarrow nrow(train_{user})$ 
   $n\_ventanas\_test \leftarrow nrow(test_{user})$ 

  for  $i \leftarrow 1$  to  $n\_ventanas\_test$  do
     $scores\_auxiliares \leftarrow vector()$ 

    for  $j \leftarrow 1$  to  $n\_ventanas\_train$  do
       $salida \leftarrow distanciaEuclidea(test_{user}[i,], train_{user}[j,])$ 
       $score\_auxiliares \leftarrow c(score\_auxiliares, salida)$ 
    end
     $seleccion \leftarrow min(score\_auxiliares)$ 
     $scoreAutentico_{user} \leftarrow c(scoreAutentico_{user}, seleccion)$ 
  end
end

```

Algorithm 2: Algoritmo para obtener los scores de los usuarios impostores

Input Tabla de datos de entrenamiento de todos los usuarios $train_i$
 Tabla de datos de prueba de todos los usuarios $test_i$
 Usuarios disponibles en la forma experimental $usuariosDisponibles$
Output Scores impostores de todos los usuarios $scoreImpostores_i$

```

for  $user\_train$  in  $usuariosDisponibles$  do
  for  $user\_test$  in  $usuariosDisponibles$  do
    if  $user\_test \neq user\_train$  then
       $n\_ventanas\_train \leftarrow nrow(train_{user\_train})$ 
       $n\_ventanas\_test \leftarrow nrow(test_{user\_test})$ 

      for  $i \leftarrow 1$  to  $n\_ventanas\_test$  do
         $scores\_auxiliares \leftarrow vector()$ 

        for  $j \leftarrow 1$  to  $n\_ventanas\_train$  do
           $salida \leftarrow distanciaEuclidea(test_{user\_test}[i,], train_{user\_train}[j,])$ 
           $score\_auxiliares \leftarrow c(score\_auxiliares, salida)$ 
        end
         $seleccion \leftarrow min(score\_auxiliares)$ 
         $scoreImpostores_{user\_train} \leftarrow c(scoreImpostores_{user\_train}, seleccion)$ 
      end
    end
  end
end

```

El único parámetro por fijar en este algoritmo es el valor de k , el cual depende fundamentalmente de los datos. De manera general, valores grandes de k reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas.

Para los datos originales, sin ningún tipo de preprocesamiento, únicamente eliminando el ruido de manera manual, se han probado valores impares de k entre 1 y 25, tanto en el dominio del tiempo como de la frecuencia utilizando o el módulo o las 3 componentes XYZ juntas. Los resultados se pueden ver en los gráficos de las figuras 5.4 y 5.5 para el dominio del tiempo y 5.6 y 5.7 para el dominio de la frecuencia.

El eje X de los gráficos de la figura 5.4, 5.5, 5.6 y 5.7 indica el valor de k y el eje Y el valor de la tasa de equierror. Idealmente es mejor cuánto EER más pequeño. Aunque hay excepciones, suele ser mejor utilizar $k=1$, resultando el mejor de manera global. Otra cosa que se observó analizando cada usuario es que los usuarios que tienen comportamientos extraños y diferentes frente al resto, son aquellos que llevaban el reloj/pulsera en la mano dominante. Ejemplo de ello es el usuario 8 que obtiene resultados malos, llegando incluso en el dominio de la frecuencia con el módulo a obtener EER de siempre 0.5 (igual que aleatorio) (gráfico 5.7); lo mismo le ocurre al usuario 15 obteniendo valores de 0.5 tanto en el módulo como en XYZ del dominio de la frecuencia (gráficos 5.6 y 5.7). Y resultados que aunque no son de 0.5, si son excesivamente malos con $k=1$ ocurre en el usuario 11 para el dominio del tiempo y las 3 componentes XYZ (gráfico 5.4).

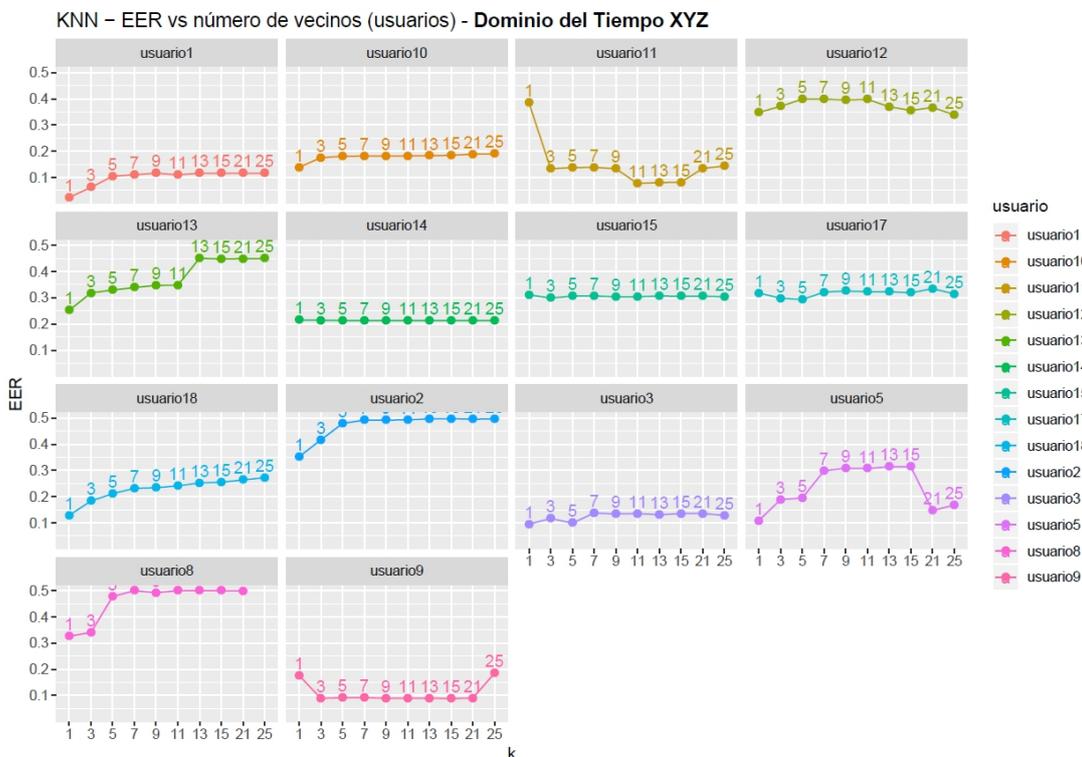


Figura 5.4: EER variando k en KNN, para cada usuario, Dominio del Tiempo XYZ.

Finalmente, por su sencillez y rapidez, ya que necesita menor tiempo de cómputo, y por rendimiento se ha seleccionado el valor impar de $k=1$, llamando en este caso al algoritmo como *Nearest*

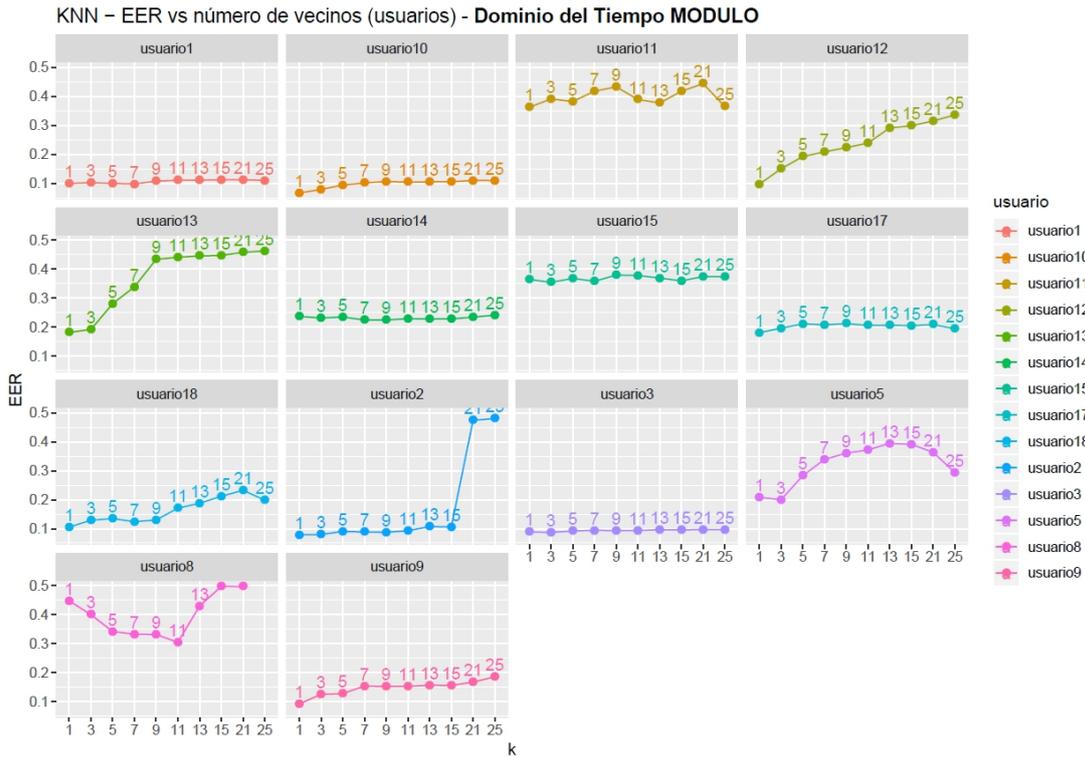


Figura 5.5: EER variando k en KNN, para cada usuario, Dominio del Tiempo Módulo.

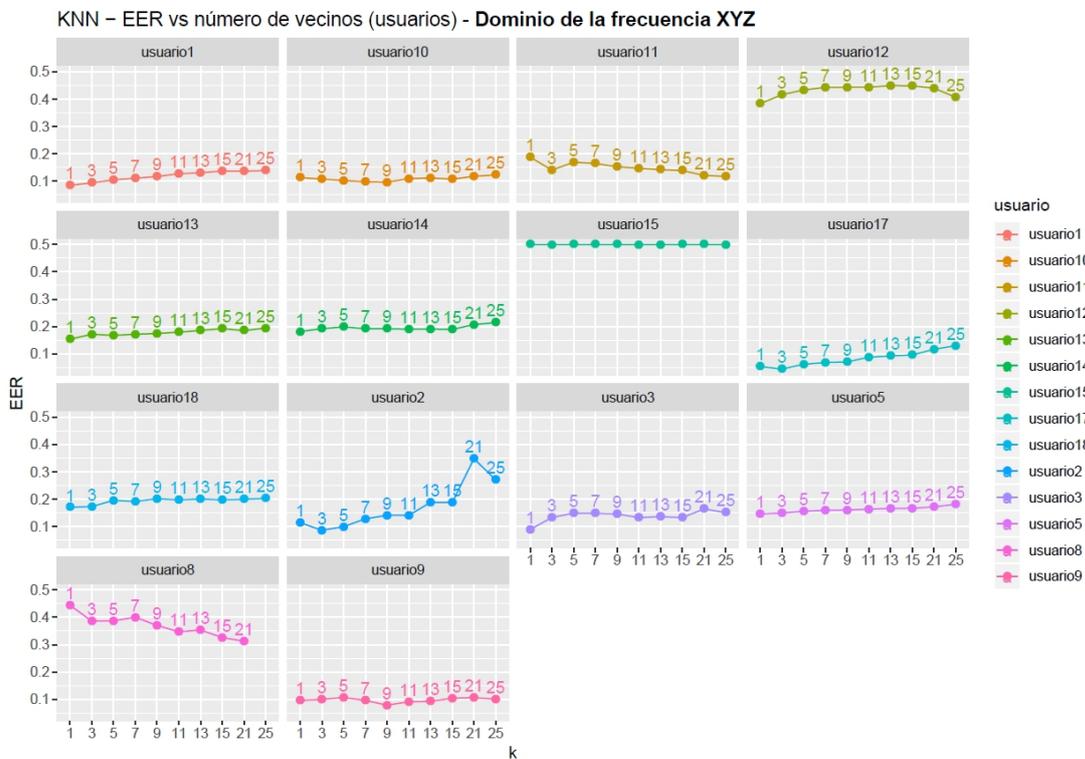


Figura 5.6: EER variando k en KNN, para cada usuario, Dominio de la Frecuencia XYZ.

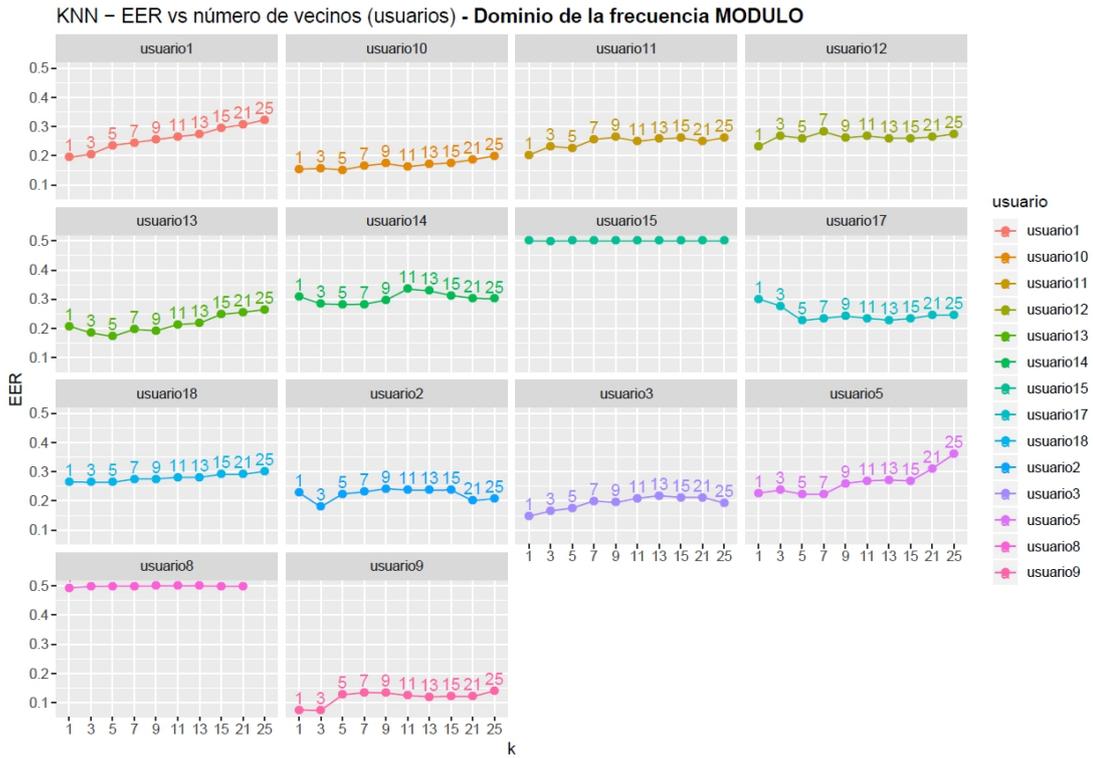


Figura 5.7: EER variando k en KNN, para cada usuario, Dominio de la Frecuencia Módulo.

Neighbor Algorithm o algoritmo del vecino más cercano.

Capítulo 6

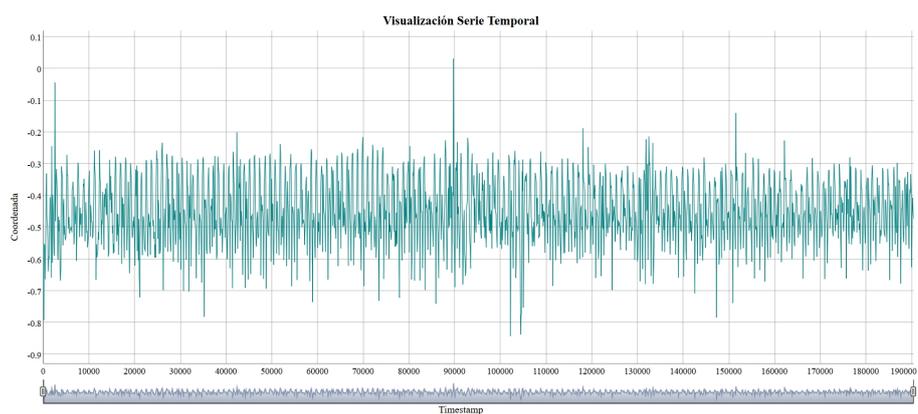
Experimentos: Resultados finales

Utilizando todo el trabajo realizado en el estudio prospectivo, donde se había decidido utilizar el dispositivo de la pulsera *Microsoft* por ser un entorno más cerrado que el del reloj *Motorola*, permitiendo la adquisición de los datos a través de su SDK, con mayor fiabilidad y mejor conexión para tomar las mejores decisiones. Como sensor el acelerómetro por ser el más habitualmente usado en la bibliografía. Consiguiendo tras las decisiones de preprocesamiento y el análisis de los parámetros la optimización de un sistema final. Si se quiere conocer todo el estudio realizado que aquí se menciona, se puede encontrar en otro proyecto del Grado en Estadística [5].

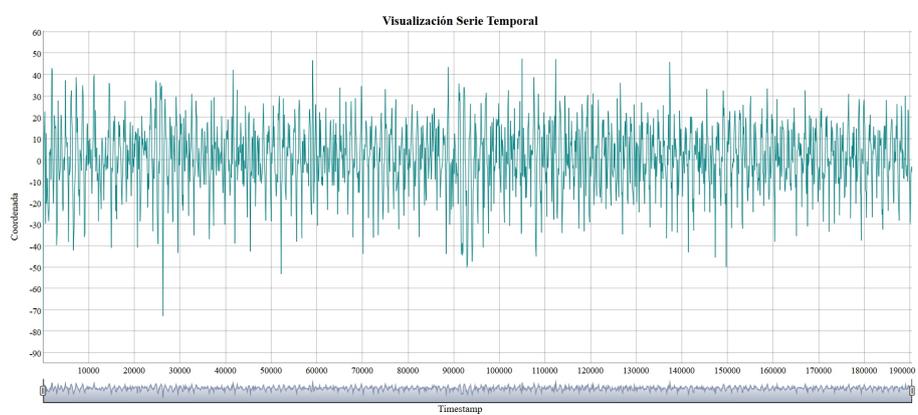
Se tienen disponibles los datos de otro sensor, el giroscopio, cuya señal es muy parecida a la del acelerómetro, puede verse en los gráficos de la figura 6.1, donde se representa al mismo usuario en la misma muestra y componente. Además, en el análisis de autocorrelaciones realizado en otro TFG [4] anterior del Grado en Ingeniería Informática por Daniel González Alonso, tabla 6.1, se puede ver que las autocorrelaciones en ambos sensores e incluso dispositivos son muy parecidas, consiguiéndose los valores más altos con tamaños de ventana, τ , iguales o muy similares. Se puede concluir que ambos sensores son muy parecidos en autocorrelación y forma, pudiendo utilizar las mismas decisiones en los dos.

τ	Autocorrelaciones											
	Micro						Moto					
	ACC			GYR			ACC			GYR		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
500	0,19	-0,83	-0,50	-0,61	-0,82	-0,83	0,33	-0,73	-0,28	-0,64	-0,68	-0,74
583	-0,13	-0,69	-0,45	-0,56	-0,68	-0,69	0,09	-0,62	-0,22	-0,55	-0,51	-0,62
666	-0,38	-0,40	-0,29	-0,38	-0,40	-0,42	-0,30	-0,39	-0,10	-0,38	-0,23	-0,39
749	-0,39	-0,03	-0,09	-0,11	-0,07	-0,07	-0,50	-0,11	-0,04	-0,18	-0,03	-0,14
832	-0,15	0,35	0,18	0,21	0,29	0,31	-0,35	0,18	0,00	0,06	0,12	0,13
915	0,29	0,67	0,49	0,53	0,65	0,67	0,12	0,49	0,20	0,38	0,37	0,48
998	0,71	0,84	0,71	0,77	0,88	0,90	0,64	0,72	0,42	0,70	0,67	0,81
1081	0,72	0,83	0,71	0,76	0,86	0,88	0,76	0,77	0,44	0,81	0,73	0,89
1164	0,26	0,62	0,47	0,48	0,61	0,62	0,41	0,63	0,26	0,61	0,48	0,65
1247	-0,17	0,29	0,15	0,18	0,25	0,24	-0,12	0,35	0,07	0,29	0,14	0,27
1330	-0,38	-0,09	-0,12	-0,11	-0,11	-0,13	-0,49	0,05	-0,07	0,01	-0,08	-0,05
1413	-0,36	-0,45	-0,33	-0,38	-0,43	-0,46	-0,47	-0,23	-0,10	-0,22	-0,23	-0,30
1496	-0,11	-0,70	-0,47	-0,55	-0,67	-0,70	-0,17	-0,50	-0,22	-0,43	-0,42	-0,51

Tabla 6.1: Autocorrelaciones obtenidas para la muestra 1 y sesion 1 del usuario19



(a) Microsoft Acelerómetro



(b) Microsoft Giroscopio

Figura 6.1: Comparación de la forma de la serie temporal en los dos sensores.

En este capítulo se va a trabajar con ambos dominios, del tiempo y de la frecuencia, probando las siguientes opciones (lista 1):

- En el **dominio del tiempo (DT)**: componenteX (DT_X), componenteY (DT_Y), componenteZ (DT_Z) y modulo_DT.
- En el **dominio de la frecuencia (DF)**: componenteX (DF_X), componenteY (DF_Y), componenteZ (DF_Z) y modulo_DF.

Probando en cada caso (lista 2):

- Ambos **dispositivos**: pulsera *Microsoft* (Micro) y reloj *Motorola* (Moto).
- Ambos **tipos de sensores** en cada dispositivo: acelerómetro (ACC) y giroscopio (GYR).
- Los tres **procedimientos experimentales** en cada combinación dispositivo/sensor: Monosesión-Monomuestra (MonoMono), Multisesión-Monomuestra (MultiMono) y Multisesión-Multimuestra (MultiMulti).

Se tendrá, por tanto, una tabla para cada opción a probar de la lista 1, en total 8 tablas, cada una de ellas con 12 columnas (cada experimento de la lista 2). El esquema de cada tabla se puede ver en la figura 6.2.

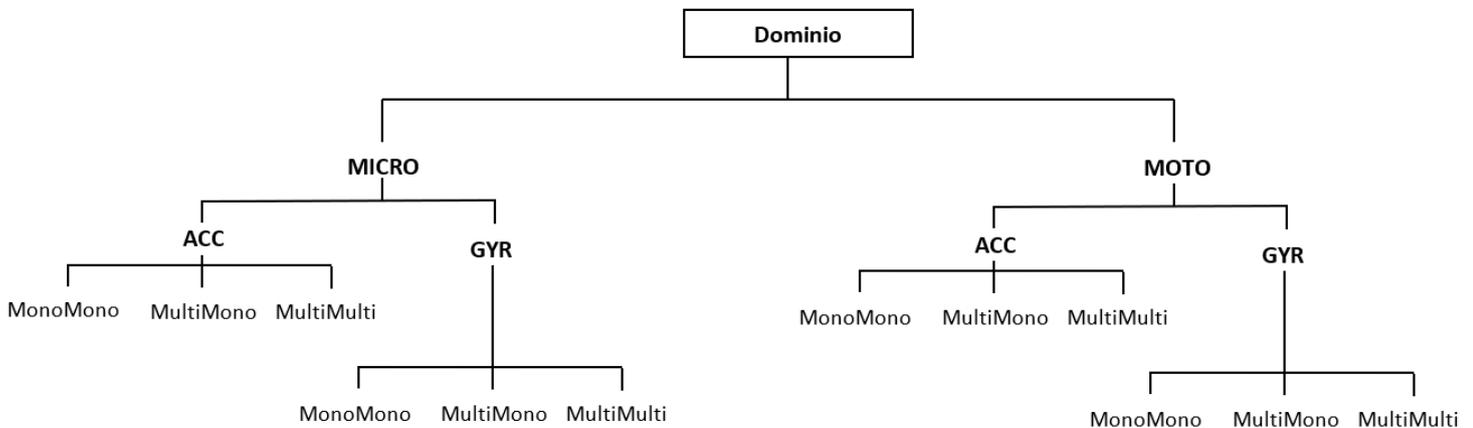


Figura 6.2: Esquema tablas del Rendimiento Final

Sobre cada columna de cada tabla, se aplicarán todas las decisiones tomadas en el TFG que arriba se menciona [5]. De manera resumida son:

- Utilizar los datos originales, realizando la **segmentación automática** basada en la autocorrelación con el criterio de eliminar las ventanas con valor de autocorrelación en valor absoluto inferior a $\frac{3}{4}$ del máximo valor de cada usuario/sesión/muestra/componente. Relajando el criterio, reduciendo $\frac{3}{4}$ en 5 centésimas, en caso de perder usuarios por insuficiente número de ventanas.

- Respecto a **preprocesamiento**, utilizar los datos originales con el filtro de la media móvil de orden 3 en el dominio del tiempo y los datos interpolados sin normalizar las amplitudes del Análisis de Fourier en el dominio de la frecuencia.
- Utilizar **tamaños de ventana** entre 6 y 10 ciclos.
- Utilizar la **técnica de fusión** sobre los scores obtenidos con el clasificador con el estadístico de la mediana, tamaño de ventana para la fusión 4 y solapamiento 2.
- Para llegar al **rendimiento final**, volver a aplicar la fusión para los tamaños de ventana seleccionados, entre 6 y 10 utilizando una secuencia equiespaciada de valores para conseguir la misma longitud en todos los tamaños y poder aplicar, después, el estadístico de la media, siguiendo lo explicado en [5].

Como resultado, se va a obtener la tasa de equierror de cada usuario y el valor medio de todos ellos. El objetivo aquí va a ser, en primer lugar, comparar los resultados de ambos sensores (ACC y GYR). Obtener resultados similares significaría que ambos sensores son parecidos y podría ser apropiado usarlos de manera complementaria, en cambio si los resultados son malos, no sé sabrían las razones y habría que repetir y cambiar el preprocesamiento y el análisis de los parámetros para el giroscopio, ya que hemos empleado las conclusiones del experimento realizado sobre el acelerómetro. De la misma manera, comparar los resultados de los dos dispositivos (Micro y Moto) para ver si se pueden extrapolar las decisiones a “cualquier” dispositivo comercial. Si los resultados en ambos dispositivos fueran similares, sería indicativo de la independencia del sistema de reconocimiento con respecto al dispositivo, es decir, si un usuario cambia de dispositivo ponible, no hay que cambiar el sistema de reconocimiento. Utilizando en cada combinación dispositivo/-sensor los 3 procedimientos experimentales: *Monosesión-Monomuestra* como prueba inicial y caso más favorable por usar muestras tomadas en la misma sesión, *Multisesión-Monomuestra* para ver hasta dónde puede llegar nuestro sistema, probando la variabilidad del rasgo biométrico con el tiempo y *Multisesión-Multimuestra* como escenario intermedio utilizando muestras de distintas sesiones para intentar paliar la variabilidad del rasgo biométrico con el tiempo, como se explicó en el apartado 5.3. Para este primer objetivo de comparar sensores/dispositivos y procedimientos experimentales se va a utilizar el error medio de todos los usuarios disponibles y su visualización gráfica, para poder compararlo de manera tanto numérica como visual.

El número total de usuarios disponibles en *Monosesión-Monomuestra* y *Multisesión-Multimuestra* son los mismos, 14, debido a las deficiencias de los datos de que disponemos y a que no todos los usuarios tienen todas las sesiones y dos muestras en cada sesión. En cambio, en *Multisesión-Monomuestra* se tienen 18 usuarios por utilizar como *prueba auténtica* la primera muestra de la segunda sesión, ya que la mayoría de los usuarios, no tienen dos muestras por sesión, pero sí dos sesiones con una muestra en cada una de ellas.

El segundo objetivo, aprovechando que se tienen pocos usuarios, es hacer un estudio detallado de cada uno de ellos, intentando ver cuáles son mejores o peores y las razones que lo causan. Para

este estudio, se va a utilizar el módulo de los datos, por ser una buena alternativa al uso de las 3 componentes y haberse visto que funciona de manera muy parecida o mejor.

Por último, como tercer objetivo, se va a estudiar si de verdad merece la pena utilizar ambos sensores de manera complementarias, y qué mejoras se podrían llegar a conseguir por este camino. En este caso, se va a trabajar con *Multisesión-Multimuestra* por ser el escenario más realista y apropiado. En este estudio, también se van a observar los resultados de la métrica (EER) para cada usuario, a modo de comparativa, para ver si los usuarios malos en un sensor mejoran con el otro.

A continuación, se muestran los resultados para cada dominio por separado.

6.1. Dominio del tiempo

El dominio del tiempo es el más ampliamente utilizado en la bibliografía, siendo también el dominio en el que se capturan los datos.

En las tablas 6.2 y 6.3 se muestran las tasas de equierror medias, visualizadas gráficamente en las figuras 6.3 y 6.4. Lo más normal sería que en *Monosesión-Monomuestra* se obtuviera el mejor resultado (EER), bastante similar al procedimiento experimental *Multisesión-Multimuestra*, pero que los resultados empeorasen en *Multisesión-Monomuestra*. Esto ocurre en la componente X del acelerómetro en MICRO, la componente Y, Z y el módulo en el giroscopio de MICRO y el acelerómetro de MOTO y en la componente X, Z y el módulo del giroscopio de MOTO. Sin embargo, en otros casos no se cumple, como la componente X del giroscopio en MICRO donde *Multisesión-Monomuestra* funciona bien. La causa puede ser la variabilidad de los datos entre sesiones donde la sesión 2 podría ser bastante peor que la 1, que es la que se utiliza en *Monosesión-Monomuestra*, pero tampoco se tienen suficientes datos para extraer más conclusiones. Esto también se puede ver en los gráficos de las figuras 6.3 y 6.4, correspondientes al dispositivo de la pulsera *Microsoft* y al reloj *Motorola* respectivamente, donde se ve que la mayoría de las veces, como era de esperar, la línea verde (Multisesión Monomuestra) está por encima de las naranja y azul, pero no siempre.

Por otro lado, se puede ver como los resultados son muy parecidos entre sensores y dispositivos. Si utilizamos únicamente el escenario de *Multisesión-Multimuestra*, la mejor tasa de equierror está en torno a los mismos valores, con diferencias no superiores al 2% en valor absoluto (tabla 6.4). En la pulsera de *Microsoft* los mejores resultados se consiguen con el módulo y en el reloj *Motorola* con la componente X y la Y para el acelerómetro y el giroscopio, respectivamente, con diferencias de un 1% e inferiores al 3% con respecto a usar el módulo.

Visto el rendimiento medio, pasamos al análisis detallado por individuo de la base de datos. De las tablas anteriores, podemos ver como no hay una componente que sea superior al resto tanto para acelerómetro como para giróscopo, siendo la opción que presenta, en general, mejores resultados el módulo. Por esta razón, este análisis lo vamos a realizar solo para ese caso, ya que si

	MICRO					
	ACC			GYR		
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
DT_X	0.1196	0.1554	0.1347	0.1507	0.1311	0.1483
DT_Y	0.1173	0.1886	0.1804	0.0991	0.1691	0.1299
DT_Z	0.1046	0.0989	0.129	0.1792	0.1859	0.171
modulo_DT	0.1008	0.0833	0.1241	0.1164	0.1635	0.1026

Tabla 6.2: Resultados EER medio del Dominio del Tiempo (Dispositivo MICRO)

	MOTO					
	ACC			GYR		
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
DT_X	0.181	0.2206	0.1142	0.1687	0.2776	0.1801
DT_Y	0.1024	0.1632	0.1335	0.1326	0.128	0.124
DT_Z	0.157	0.1918	0.1336	0.1205	0.2444	0.1307
modulo_DT	0.1277	0.1545	0.1217	0.1307	0.2071	0.1508

Tabla 6.3: Resultados EER medio del Dominio del Tiempo (Dispositivo MOTO)

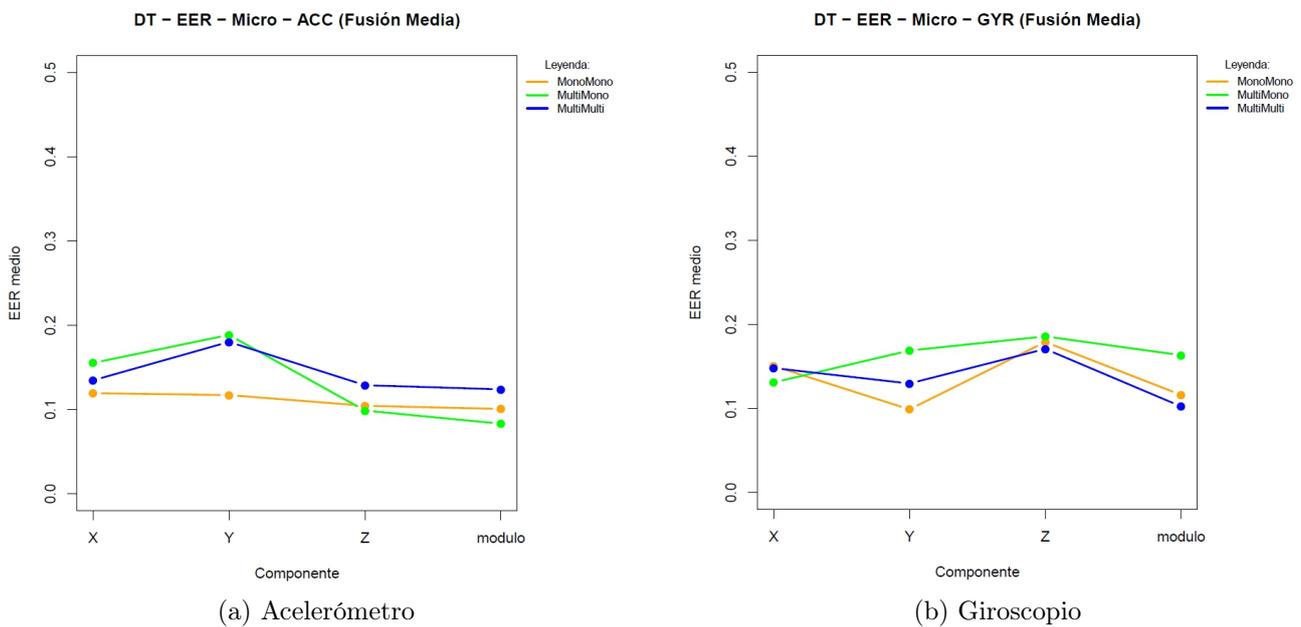


Figura 6.3: Resultados EER medio - Dominio del Tiempo (Dispositivo MICRO).

Dominio del tiempo	ACC	GYR
MICRO	0.1241	0.1026
MOTO	0.1142	0.124

Tabla 6.4: Resumen Mejores Resultados en Multisesión-Multimuestra - dominio del tiempo.

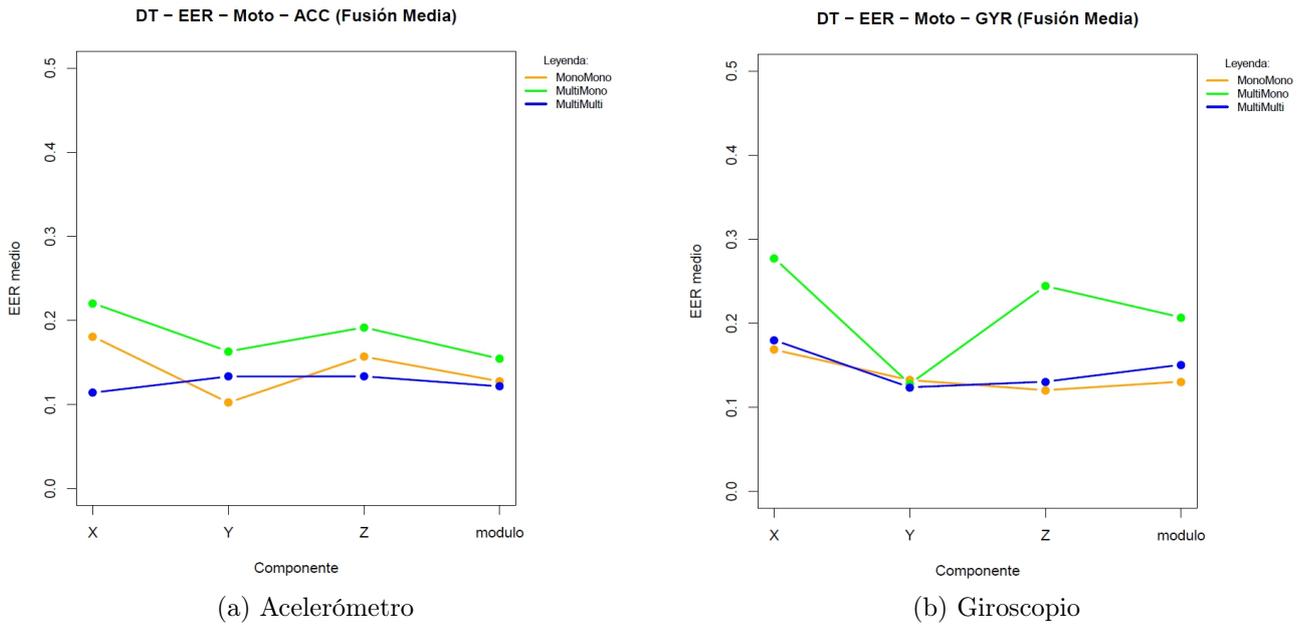


Figura 6.4: Resultados EER medio - Dominio del Tiempo (Dispositivo MOTO).

no la gran cantidad de resultados haría imposible extraer conclusiones. Los resultados se pueden ver en las tablas 6.5 y 6.6. Analizando cada usuario, del que se tiene información en todos los procedimientos experimentales, se puede concluir la siguiente categorización del acelerómetro de la pulsera *Microsoft*.

- Usuarios que empeoran al introducir los datos de las dos sesiones: 1, 9
- Al contrario, usuarios que mejoran al introducir los datos de las dos sesiones: 8, 12
- Usuarios con resultados razonablemente buenos (peor EER en MultiMono y mejor en MultiMulti): 2, 13, 14
- Usuarios con resultados raros, ya que en MultiMono consiguen buena tasa de error, pero en MultiMulti y MonoMono no tan buena: 3, 11
- Usuarios buenos en cualquier procedimiento experimental: 10, 18
- Usuarios malos en cualquier procedimiento experimental: 15

Al cambiar de sensor en la pulsera de *Microsoft*, el usuario 2, 9 y 11 pasan a tener resultados razonablemente buenos en todos los procedimientos experimentales, al contrario que el 10 y el 14 que pasan a tener resultados malos en todos los procedimientos experimentales. El usuario 18 también pasa a ser malo en todos los procedimientos experimentales, aunque mejora al introducir los datos de las dos sesiones (MultiMulti). Estas conclusiones pueden llevar a pensar lo que ya se sospechaba, que una combinación de ambos sensores podría ser apropiada.

Al cambiar de dispositivo, al reloj *Motorola*, el estudio de los usuarios cambia. Para el acelerómetro se puede decir lo siguiente.

- Usuarios que empeoran al introducir los datos de las dos sesiones: 2, 10, 18
- Al contrario, usuarios que mejoran al introducir los datos de las dos sesiones: 11
- Usuarios con resultados razonablemente buenos (peor EER en MultiMono, pero bueno en MultiMulti): 1, 12
- Usuarios buenos en cualquier procedimiento experimental: 3, 8, 9, 14
- Usuarios malos en cualquier procedimiento experimental: 13, 15

Al cambiar de sensor en el reloj de *Motorola*, el usuario 2 mejora aunque se sigue cumpliendo el hecho de que empeore al introducir los datos de las dos sesiones. El usuario 10 continúa siendo un usuario malo, no empeorando al introducir los datos de las 2 sesiones, mientras que el usuario 3, sí empeora al introducir los datos de las dos sesiones. El usuario 11 pasa a ser un usuario bueno en cualquier procedimiento experimental, mientras que el 13 que era un usuario malo, mejora en MultiMulti y los usuarios 8, 9, 12 y 14 pasan a ser usuarios malos en cualquier procedimiento experimental. Viéndose de nuevo, como una combinación de ambos sensores podría ser apropiada.

La razón por la que la categorización de los usuarios es distinta en los 2 dispositivos se debe al comportamiento variable del ser humano, pudiéndose atribuir a muchos factores como la colocación de la pulsera/reloj, a que una de ellas les haya resultado más cómoda o simplemente que se la haya sujeto mejor o peor a la muñeca. No habiéndose encontrado ninguna explicación en la visualización de las series de tiempo ni en el análisis de las autocorrelaciones.

Por último, y para hacer un estudio inicial del uso de ambos sensores, se va a utilizar el procedimiento experimental *Multisesión-Multimuestra*, como ya se ha explicado en la introducción de este capítulo. Los resultados del módulo se muestran en las tablas referenciadas como 6.7, en las cuales se han eliminado aquellos usuarios de los que no se tiene información en este procedimiento experimental. En la pulsera de Microsoft, aunque siguen existiendo usuarios con EER altos, se consiguen mejoras en muchos otros, como el 5, 8, 9, 10, 11, 12, 14 o 17, llegando a conseguir un resultado final bastante bueno de prácticamente el 6% de tasa de equierror, mucho mejor que lo que se tenía utilizando solamente uno de los dos sensores. Respecto al reloj de *Motorola*, ocurre lo mismo, consiguiendo un EER final usando ambos sensores del 7.6%. En ambas tablas, se puede ver, en la parte inferior, como en ambos casos prácticamente la mitad de las veces funciona mejor el acelerómetro y la otra mitad el giroscopio, mostrando como ambos sensores se pueden usar de manera complementaria.

Los resultados de este mismo estudio en las componentes X, Y y Z se pueden ver en las tablas 6.8, 6.9 y 6.10, respectivamente. En todos los casos, la tasa de equierror media usando ambos sensores mejora con respecto a utilizar únicamente uno de los dos, resultando usuarios mejores tanto en el acelerómetro como en el giroscopio, indistintamente. Para verlo más rápido, si el EER del usuario se encuentra marcado con color azul claro es que funciona mejor el acelerómetro, mientras que si está marcado con azul más oscuro funciona mejor el giroscopio. Por otro lado, los

Dominio del tiempo - modulo						
Micro						
ACC			GYR			
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
usuario 1	0	0	0.1722	0	0.2405	0.22
usuario 2	0.04	0.2602	0.0679	0.0694	0.0882	0.0909
usuario 3	0.1579	0	0.1109	0.1221	0	0.0743
usuario 4	NA	0	NA	NA	0	NA
usuario 5	0.1812	NA	0.1604	0.0606	NA	0.0329
usuario 6	NA	0.0594	NA	NA	0.1282	NA
usuario 7	NA	0	NA	NA	0.3736	NA
usuario 8	0.2658	0.3203	0.1306	0.3942	0.4615	0.0087
usuario 9	0.0367	0.0865	0.2791	0.0429	0.0588	0.0084
usuario 10	0	0	0.0059	0.1014	0.0882	0.2205
usuario 11	0.2125	0	0.139	0.0152	0.0149	0.026
usuario 12	0.1297	0.0109	0.0579	0.2169	0.0972	0.1691
usuario 13	0.0106	0.0664	0.0717	0	0.1905	0.0084
usuario 14	0.0323	0.0838	0.0544	0.1351	0.1795	0.1234
usuario 15	0.1868	0.415	0.2999	0.2063	0.4175	0.3914
usuario 16	NA	0	NA	NA	0.05	NA
usuario 17	0.1476	NA	0.1586	0	NA	0
usuario 18	0.0106	0.0478	0.0287	0.265	0.1646	0.0629
usuario 19	NA	0.0187	NA	NA	0.1169	NA
usuario 20	NA	0.1311	NA	NA	0.2734	NA
MEDIA	0.1008	0.0833	0.1241	0.1164	0.1635	0.1026

Tabla 6.5: Tabla completa del módulo en el Dominio del Tiempo (Dispositivo MICRO)

Dominio del tiempo - modulo						
Moto						
ACC			GYR			
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
usuario 1	0	0.1282	0	0	0.2857	0
usuario 2	0.0769	0.0571	0.222	0.0517	0	0.0902
usuario 3	0	0	0.037	0.0957	0	0.1589
usuario 4	NA	0	NA	NA	0.082	NA
usuario 5	0.15	NA	0.0798	0	NA	0
usuario 6	NA	0	NA	NA	0.1688	NA
usuario 7	NA	0.3929	NA	NA	0.1129	NA
usuario 8	0	0.0155	0	0.2388	0.3855	0.1574
usuario 9	0	0.0153	0.0396	0	0.3714	0.1416
usuario 10	0.3939	0.1316	0.3488	0.3205	0.4	0.277
usuario 11	0.2411	0.0769	0.0764	0	0	0
usuario 12	0.0761	0.2975	0.0984	0.1333	0.3953	0.3214
usuario 13	0.25	0.28	0.233	0.1	0.381	0.0588
usuario 14	0	0	0.0595	0.375	0.2	0.4431
usuario 15	0.3474	0.4833	0.3196	0.0377	0.4842	0.2313
usuario 16	NA	0.0981	NA	NA	0.0508	NA
usuario 17	0.0405	NA	0.0159	0.0364	NA	0
usuario 18	0.212	0.0366	0.174	0.4409	0.037	0.2315
usuario 19	NA	0.2983	NA	NA	0.3724	NA
usuario 20	NA	0.4692	NA	NA	0	NA
MEDIA	0.1277	0.1545	0.1217	0.1307	0.2071	0.1508

Tabla 6.6: Tabla completa del módulo en el Dominio del Tiempo (Dispositivo MOTO)

	DT módulo (MICRO)		
	ACC	GYR	MEJOR
	MultiMulti	MultiMulti	
1	0.1722	0.22	0.1722
2	0.0679	0.0909	0.0679
3	0.1109	0.0743	0.0743
5	0.1604	0.0329	0.0329
8	0.1306	0.0087	0.0087
9	0.2791	0.0084	0.0084
10	0.0059	0.2205	0.0059
11	0.139	0.026	0.026
12	0.0579	0.1691	0.0579
13	0.0717	0.0084	0.0084
14	0.0544	0.1234	0.0544
15	0.2999	0.3914	0.2999
17	0.1586	0	0
18	0.0287	0.0629	0.0287
MEDIA	0.1241	0.1026	0.0604
	7	7	

(a) Pulsera Microsoft

	DT módulo (MOTO)		
	ACC	GYR	MEJOR
	MultiMulti	MultiMulti	
1	0	0	0
2	0.222	0.0902	0.0902
3	0.037	0.1589	0.037
5	0.0798	0	0
8	0	0.1574	0
9	0.0396	0.1416	0.0396
10	0.3488	0.277	0.277
11	0.0764	0	0
12	0.0984	0.3214	0.0984
13	0.233	0.0588	0.0588
14	0.0595	0.4431	0.0595
15	0.3196	0.2313	0.2313
17	0.0159	0	0
18	0.174	0.2315	0.174
MEDIA	0.1217	0.1508	0.0761
	6.5	7.5	

(b) Reloj Motorola

Tabla 6.7: Combinación ACC/GYR del módulo en el Dominio del Tiempo.

resultados en ambos dispositivos son muy similares, y aunque se tienen pocos usuarios y no se cumple que siempre sean buenos y malos los mismos, se complementan consiguiendo resultados del rendimiento final muy parecidos.

En cada tabla se han marcado con negrita las tasas de equierror más altas, generalmente, superiores al 10% y si se comparan los usuarios marcados, que serían los “malos”, se puede ver cómo van cambiando, no siguiendo ningún patrón concreto, pero sí tendiendo a repetirse los mismos usuarios dentro del mismo dispositivo. Se indican los usuarios malos en la tabla 6.11. Encontrándose aquí la mayor dificultad del problema que se está resolviendo, y en general de cualquier biometría, ya que está demostrado que el comportamiento humano es variable y puede verse afectado por muchos factores.

Componente	Usuarios con mayor EER MICRO	Usuarios con mayor EER MOTO
X	1, 2, 3, 9, 11, 15	1, 8, 10, 15, 18
Y	2, 3, 12, 14, 15	12, 13, 15, 18
Z	1, 3, 8, 9, 12, 15, 17	8, 10, 12, 18
módulo	1, 15	10, 15, 18

Tabla 6.11: Resumen Usuarios EER alto Combinación ACC/GYR - dominio del tiempo.

	DT-X (MICRO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.1886	0.2771	0.1886
2	0.128	0.1259	0.1259
3	0.1227	0.1354	0.1227
5	0.0063	0.0544	0.0063
8	0.1257	0.0473	0.0473
9	0.3902	0.1361	0.1361
10	0.0342	0.1808	0.0342
11	0.1556	0.4206	0.1556
12	0.0667	0.2266	0.0667
13	0.3274	0	0
14	0.0746	0.2029	0.0746
15	0.1961	0.2081	0.1961
17	0.0703	0.031	0.031
18	0	0.0296	0
MEDIA	0.1347	0.1483	0.0847
	9	5	

(a) Pulsera Microsoft

	DT-X (MOTO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.119	0.1154	0.1154
2	0.0271	0.3252	0.0271
3	0	0.2786	0
5	0.0268	0.0873	0.0268
8	0.1837	0.1301	0.1301
9	0.1308	0	0
10	0.3762	0.2365	0.2365
11	0.0441	0.4119	0.0441
12	0.0493	0.1704	0.0493
13	0.2199	0.0877	0.0877
14	0.0932	0.0367	0.0367
15	0.1129	0.1613	0.1129
17	0.063	0.2121	0.063
18	0.1525	0.2679	0.1525
MEDIA	0.1142	0.1801	0.0773
	8	6	

(b) Reloj Motorola

Tabla 6.8: Combinación ACC/GYR de la componente X en el Dominio del Tiempo.

	DT-Y (MICRO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.2694	0.0131	0.0131
2	0.4922	0.1872	0.1872
3	0.1394	0.1277	0.1277
5	0	0.2312	0
8	0.085	0.093	0.085
9	0.2429	0.1041	0.1041
10	0.0357	0.1603	0.0357
11	0.414	0.105	0.105
12	0.3229	0.2912	0.2912
13	0.0629	0	0
14	0.1391	0.2091	0.1391
15	0.2592	0.2069	0.2069
17	0.0338	0.0186	0.0186
18	0.0288	0.071	0.0288
MEDIA	0.1804	0.1299	0.0959
	5	9	

(a) Pulsera Microsoft

	DT-Y (MOTO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0	0.3289	0
2	0.1593	0.0559	0.0559
3	0.1724	0.0089	0.0089
5	0.0446	0.0724	0.0446
8	0.0671	0.181	0.0671
9	0	0.0265	0
10	0.1724	0.0586	0.0586
11	0.1086	0.0704	0.0704
12	0.3435	0.1883	0.1883
13	0.3333	0.2792	0.2792
14	0	0.0091	0
15	0.2696	0.223	0.223
17	0	0.0917	0
18	0.1977	0.1418	0.1418
MEDIA	0.1335	0.124	0.0813
	6	8	

(b) Reloj Motorola

Tabla 6.9: Combinación ACC/GYR de la componente Y en el Dominio del Tiempo.

DT-Z (MICRO)			
	ACC	GYR	
	MultiMulti	MultiMulti	MEJOR
1	0.2044	0.2834	0.2044
2	0.0907	0.076	0.076
3	0.1134	0.1272	0.1134
5	0.0934	0.1179	0.0934
8	0.1744	0.1984	0.1744
9	0.1571	0.2523	0.1571
10	0	0.1986	0
11	0.1206	0.0769	0.0769
12	0.2847	0.2209	0.2209
13	0.0065	0	0
14	0.0529	0.1767	0.0529
15	0.2986	0.3361	0.2986
17	0.1963	0.2681	0.1963
18	0.013	0.0607	0.013
MEDIA	0.129	0.171	0.1198
	10	4	

(a) Pulsera Microsoft

DT-Z (MOTO)			
	ACC	GYR	
	MultiMulti	MultiMulti	MEJOR
1	0.0663	0.1033	0.0663
2	0	0.2174	0
3	0	0.186	0
5	0	0.0071	0
8	0.3662	0.2107	0.2107
9	0.0438	0.0074	0.0074
10	0.3198	0.2147	0.2147
11	0.1725	0.0525	0.0525
12	0.3497	0.1483	0.1483
13	0.0882	0.1203	0.0882
14	0.0171	0.0938	0.0171
15	0.106	0.1092	0.106
17	0.0847	0.2278	0.0847
18	0.2563	0.131	0.131
MEDIA	0.1336	0.1307	0.0805
	8	6	

(b) Reloj Motorola

Tabla 6.10: Combinación ACC/GYR de la componente Z en el Dominio del Tiempo.

6.2. Dominio de la frecuencia

El dominio de la frecuencia es, hasta ahora, un entorno menos usado en esta biometría, pero no se ha encontrado ninguna base que diga que no se puede usar, ni nada que advierta sobre la existencia de inconvenientes para usarla. Respecto a necesidades de computo tampoco existen problemas. En el estudio prospectivo se utilizó y generaba peores resultados que el dominio del tiempo. No obstante, ahora se va a realizar un estudio más profundo, utilizando los diferentes sensores, dispositivos y escenarios experimentales.

De la misma manera que antes, en las tablas 6.12 y 6.13 se muestran las tasas de equierror medias en cada procedimiento experimental, sensor y dominio, separado por dispositivos y en los gráficos de las figuras 6.5 y 6.6 la misma información de manera gráfica. El acelerómetro de la pulsera *Microsoft* tiene un comportamiento razonable en sus componentes X, Z y el módulo, consiguiendo los mejores resultados en los escenarios de *Monosesión-Monomuestra* y *Multisesión-Multimuestra* y los peores en *Multisesión-Monomuestra*, como cabía esperar, mientras que cuando ocurren cosas menos razonables, en que *Multisesión-Monomuestra* consigue el mejor resultado, en la componente Y, las diferencias de EER medio en los tres procedimientos experimentales son pequeñas. Estas conclusiones son las mismas que se extraen del giroscopio en ambos dispositivos, habiendo sólo un caso, que es la componente Z en el acelerómetro del reloj *Motorola* en que *Multisesión-Multimuestra* consigue el mejor resultado, bastante razonable, pero *Monosesión-*

Monomuestra el peor con diferencias bastante grandes. Posiblemente las razones se encuentren en que los datos de la sesión 1, usada en *Monosesión-Monomuestra* sean bastante peores en esa componente, que los datos de la sesión 2. No obstante, se tienen pocos usuarios para extraer más conclusiones y conocer por qué ocurre esto. Respecto a los mejores resultados utilizando únicamente el escenario de *Multisesión-Multimuestra*, que pueden verse en la tabla 6.14, las tasas de equierror se encuentran en torno a los mismos valores dentro del mismo sensor, funcionando aparentemente mejor el acelerómetro que el giroscopio, con diferencias de hasta un 5%. En la pulsera de *Microsoft* los mejores resultados se consiguen con el módulo, igual que ocurría en el Dominio del Tiempo, mientras que en el reloj *Motorola* se consiguen en ambos casos con la componente X. Las diferencias del mismo sensor en distintos dispositivos son de un 2% y 1% para el acelerómetro y el giroscopio, respectivamente.

	MICRO					
	ACC			GYR		
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
DF_X	0.1006	0.1739	0.1211	0.1907	0.1804	0.1791
DF_Y	0.1877	0.1761	0.2018	0.0986	0.2008	0.179
DF_Z	0.1614	0.2245	0.1475	0.2339	0.2598	0.2413
modulo_DF	0.1669	0.2237	0.1175	0.141	0.1831	0.1681

Tabla 6.12: Resultados EER medio del Dominio de la Frecuencia (Dispositivo MICRO)

	MOTO					
	ACC			GYR		
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
DF_X	0.1376	0.2116	0.1375	0.1014	0.2081	0.1531
DF_Y	0.1817	0.2772	0.1699	0.1662	0.1549	0.1789
DF_Z	0.2293	0.178	0.1552	0.168	0.305	0.2146
modulo_DF	0.1414	0.2175	0.1549	0.1308	0.27	0.1857

Tabla 6.13: Resultados EER medio del Dominio de la Frecuencia (Dispositivo MOTO)

Dominio de la Frecuencia	ACC	GYR
MICRO	0.1175	0.1681
MOTO	0.1375	0.1531

Tabla 6.14: Resumen Mejores Resultados en Multisesión-Multimuestra - dominio de la frecuencia.

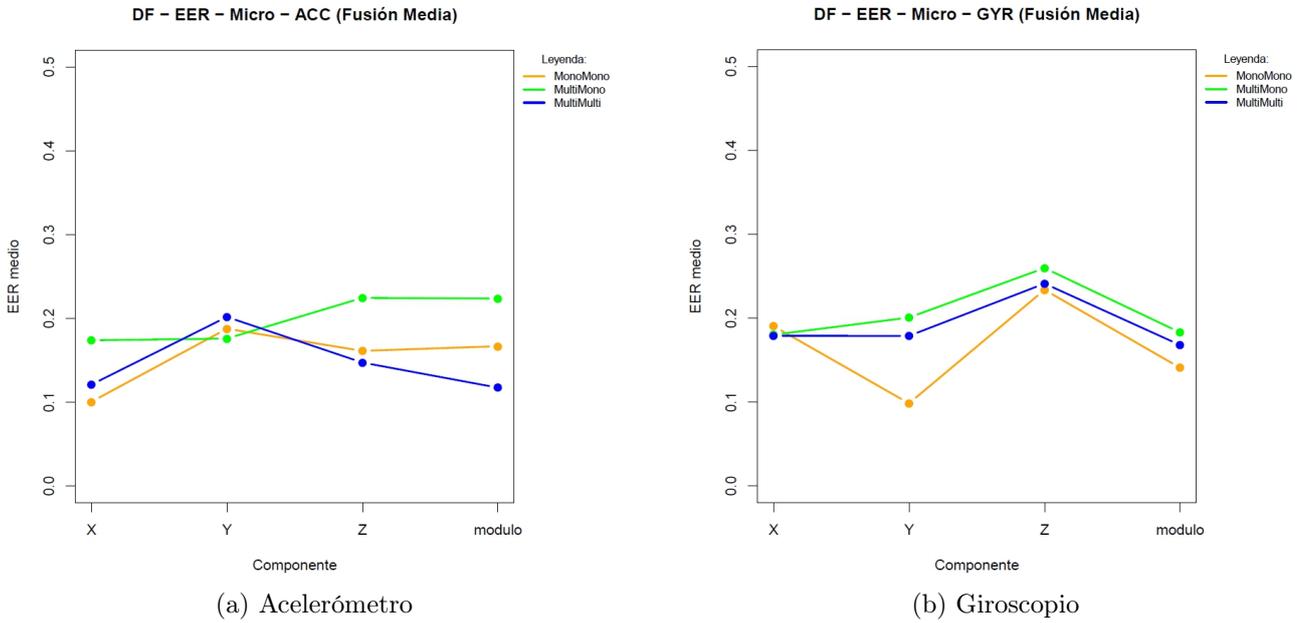


Figura 6.5: Resultados EER medio - Dominio de la Frecuencia (Dispositivo MICRO).

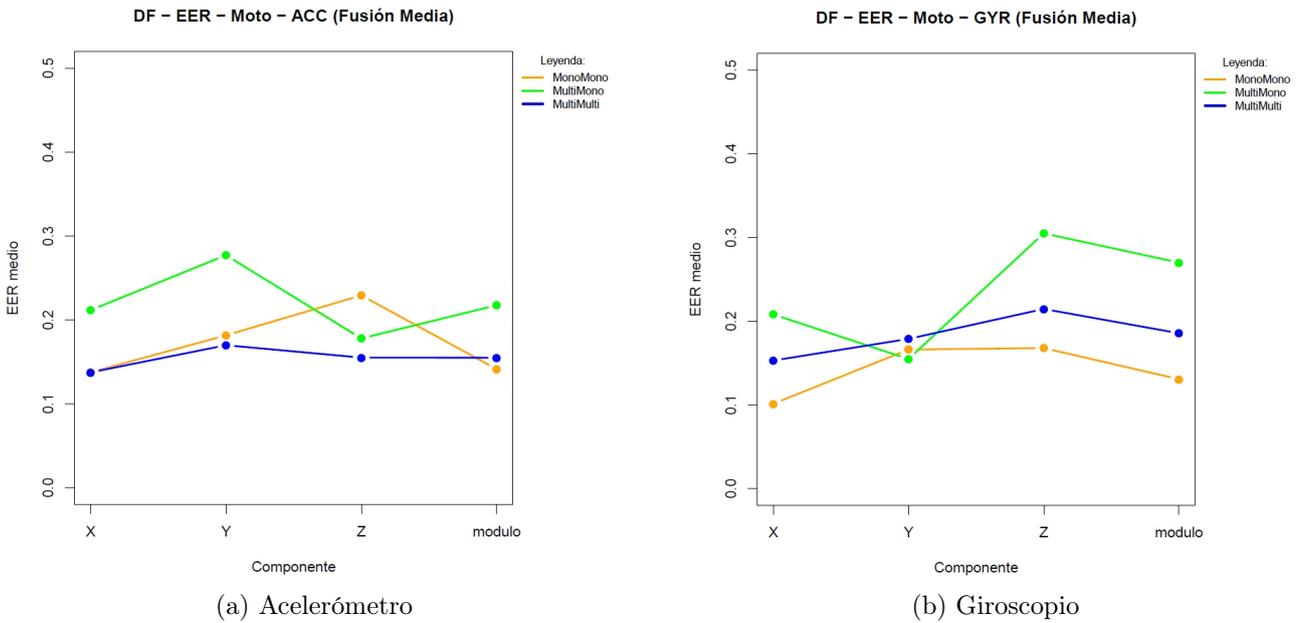


Figura 6.6: Resultados EER medio - Dominio de la Frecuencia (Dispositivo MOTO).

A continuación, se van a mostrar las tablas completas del rendimiento final separadas por dispositivos, para el módulo, e igual que antes, se va a intentar analizar cada usuario, buscando una categorización. Las tablas son la 6.15 y 6.16 y en primer lugar, los resultados del acelerómetro en la pulsera *Microsoft* son los siguientes.

- Usuarios que empeoran al introducir los datos de las dos sesiones: 9, 10, 13
- Al contrario, usuarios que mejoran al introducir los datos de las dos sesiones: 12
- Usuarios con resultados razonablemente buenos (peor EER en MultiMono, pero bueno en MultiMulti): 2, 11, 18
- Usuarios buenos en cualquier procedimiento experimental: 3
- Usuarios malos en cualquier procedimiento experimental: 1, 8, 14, 15

Al cambiar de sensor en la pulsera de *Microsoft*, el usuario 1 pasa de ser malo a conseguir resultados razonablemente buenos cuando se introducen los datos de las dos sesiones (MultiMulti). Al contrario, el usuario 3 que es bueno y el 12 empeoran cuando se introducen los datos de las dos sesiones. Los usuarios 2 y 18 pasan a ser malos en todos los procedimientos experimentales, mientras que el 9 y el 11 pasan a ser buenos y el 13 a ser razonablemente bueno, con peor EER en MultiMono, pero bueno en MonoMono y en MultiMulti.

Al utilizar el reloj *Motorola*, igual que ocurría antes, el estudio de los usuarios cambia, siendo para el acelerómetro:

- Usuarios que empeoran al introducir los datos de las dos sesiones: 9, 11, 13
- Al contrario, usuarios que mejoran al introducir los datos de las dos sesiones: 14, 18
- Usuarios con resultados razonablemente buenos (peor EER en MultiMono, pero bueno en MultiMulti): 1, 2, 12
- Usuarios malos en cualquier procedimiento experimental: 3, 8, 10, 15

Respecto a los cambios en el giroscopio del reloj *Motorola*, los usuarios malos siguen siendo malos. El usuario 2 pasa de ser razonablemente bueno a empeorar cuando se introducen los datos de las dos sesiones y el 8, 12 y 13 a ser malos en cualquier procedimiento experimental, mientras que el 11 y el 18 pasan a ser buenos en cualquier procedimiento experimental. Y el usuario 14 pasa de mejorar al introducir los datos de las dos sesiones a empeorar cuando estos se introducen. Aunque, también se puede ver que, en este dominio, se mantienen más las conclusiones entre dispositivos, coincidiendo en que los usuarios 8 y 15 son malos, el 9 y el 13 empeoran al introducir los datos de las dos sesiones y el 2 contiene resultados razonablemente buenos. Pero no entre sensores, lo que sigue haciendo ver que una combinación de ambos podría ser apropiada.

Dominio de la frecuencia - modulo						
Micro						
ACC			GYR			
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
usuario 1	0.1069	0.4025	0.174	0.0449	0.4516	0.0994
usuario 2	0	0.2335	0.0794	0.1266	0.3491	0.173
usuario 3	0	0	0.0055	0.1714	0	0.2436
usuario 4	NA	0.1187	NA	NA	0	NA
usuario 5	0.0345	NA	0.0429	0.0617	NA	0.0377
usuario 6	NA	0.1212	NA	NA	0.0952	NA
usuario 7	NA	0.2127	NA	NA	0	NA
usuario 8	1	1	0.2787	0.3846	0.3793	0.3422
usuario 9	0.0833	0.0303	0.1559	0.0723	0.0506	0.0224
usuario 10	0.0465	0.1916	0.1243	0.263	0.1522	0.3052
usuario 11	0.0635	0.2425	0.0816	0.0139	0.0272	0.0233
usuario 12	0.1434	0.3056	0.0895	0.0649	0	0.1709
usuario 13	0.0541	0.0744	0.1258	0	0.3	0.0839
usuario 14	0.1923	0.1935	0.1119	0.0658	0.1341	0.1205
usuario 15	0.399	0.4457	0.2133	0.4118	0.3966	0.4413
usuario 16	NA	0.0274	NA	NA	0	NA
usuario 17	0.1301	NA	0.1566	0.0133	NA	0.0671
usuario 18	0.0833	0.1294	0.0054	0.28	0.3208	0.2229
usuario 19	NA	0.0171	NA	NA	0.3448	NA
usuario 20	NA	0.2811	NA	NA	0.2945	NA
MEDIA	0.1669	0.2237	0.1175	0.141	0.1831	0.1681

Tabla 6.15: Tabla completa del módulo en el Dominio de la Frecuencia (Dispositivo MICRO)

Dominio de la frecuencia - modulo						
Moto						
ACC			GYR			
	MonoMono	MultiMono	MultiMulti	MonoMono	MultiMono	MultiMulti
usuario 1	0.0442	0.4753	0.0997	0.0519	0.464	0.0643
usuario 2	0.0943	0.1316	0.095	0.069	0.2418	0.1513
usuario 3	0.3307	0.3111	0.211	0.1825	0.0897	0.1765
usuario 4	NA	0	NA	NA	0.0127	NA
usuario 5	0.0417	NA	0.0415	0.046	NA	0.0667
usuario 6	NA	0.427	NA	NA	0.0641	NA
usuario 7	NA	0.098	NA	NA	0.0838	NA
usuario 8	0.1149	0.1895	0.1632	0.2648	0.3465	0.2799
usuario 9	0.018	0.2366	0.1777	0.163	0.425	0.2095
usuario 10	0.1955	0.4896	0.4078	0.25	0.4646	0.3284
usuario 11	0.0609	0.0357	0.1415	0.0405	0	0.0541
usuario 12	0.0756	0.1901	0.0492	0.1566	0.3913	0.3284
usuario 13	0.0811	0.1224	0.1174	0.1847	0.4846	0.2383
usuario 14	0.1887	0.1479	0.0701	0	0.2857	0.1479
usuario 15	0.4269	0.4615	0.2669	0.3956	0.431	0.4321
usuario 16	NA	0.1422	NA	NA	0.283	NA
usuario 17	0.1706	NA	0.2316	0	NA	0.0448
usuario 18	0.1371	0.1207	0.0962	0.027	0.0833	0.0773
usuario 19	NA	0.0645	NA	NA	0.2626	NA
usuario 20	NA	0.2718	NA	NA	0.447	NA
MEDIA	0.1414	0.2175	0.1549	0.1308	0.27	0.1857

Tabla 6.16: Tabla completa del módulo en el Dominio de la Frecuencia (Dispositivo MOTO)

DF módulo (MICRO)				DF módulo (MOTO)					
		ACC	GYR			ACC	GYR		
		MultiMulti	MultiMulti	MEJOR			MultiMulti	MultiMulti	MEJOR
1		0.174	0.0994	0.0994	1	0.0997	0.0643	0.0643	
2		0.0794	0.173	0.0794	2	0.095	0.1513	0.095	
3		0.0055	0.2436	0.0055	3	0.211	0.1765	0.1765	
5		0.0429	0.0377	0.0377	5	0.0415	0.0667	0.0415	
8		0.2787	0.3422	0.2787	8	0.1632	0.2799	0.1632	
9		0.1559	0.0224	0.0224	9	0.1777	0.2095	0.1777	
10		0.1243	0.3052	0.1243	10	0.4078	0.3284	0.3284	
11		0.0816	0.0233	0.0233	11	0.1415	0.0541	0.0541	
12		0.0895	0.1709	0.0895	12	0.0492	0.3284	0.0492	
13		0.1258	0.0839	0.0839	13	0.1174	0.2383	0.1174	
14		0.1119	0.1205	0.1119	14	0.0701	0.1479	0.0701	
15		0.2133	0.4413	0.2133	15	0.2669	0.4321	0.2669	
17		0.1566	0.0671	0.0671	17	0.2316	0.0448	0.0448	
18		0.0054	0.2229	0.0054	18	0.0962	0.0773	0.0773	
MEDIA		0.1175	0.1681	0.0887	MEDIA	0.1549	0.1857	0.1233	
		8	6			8	6		

(a) Pulsera Microsoft

(b) Reloj Motorola

Tabla 6.17: Combinación ACC/GYR del módulo en el Dominio de la Frecuencia.

Por último, igual que se ha hecho en el Dominio del Tiempo, se va a mostrar un estudio inicial usando ambos sensores utilizando *Multisesión-Multimuestra* como procedimiento experimental. Los resultados se encuentran en las tablas 6.17, 6.18, 6.19 y 6.20 para el módulo y las componentes X, Y y Z respectivamente. El formato de las tablas se mantiene, indicando el número de usuarios en que el acelerómetro y el giroscopio consiguen los mejores resultados y marcando en negrita aquellas tasas de equierror más altas. Las conclusiones son las mismas que las del Dominio del Tiempo, en todos los casos, siguiendo el estudio realizado, utilizar ambos sensores parece apropiado, reduciéndose el EER en todas las componentes. Funcionando prácticamente en la mitad de los usuarios, mejor el acelerómetro y en la otra mitad, el giroscopio, aunque con una mayor tendencia a ganar el acelerómetro. Por lo que se podría decir, que combinar ambos sensores es un buen camino a valorar y seguir para conseguir buenos resultados en esta biometría.

También se ha creado la tabla 6.21 donde se indican aquellos usuarios con EER alto, es decir, con valores grandes respecto al resto y/o superiores a aproximadamente el 10%. Si se comparan, se pueden obtener aquellos usuarios que más aparecen, pero si se observa su serie temporal o su análisis de autocorrelaciones, no siguen todos ellos un patrón concreto que los haga “malos”. Por lo que está claro, que existe una variabilidad en el comportamiento humano que lo hace difícil de categorizar, pudiéndose atribuir a factores como la colocación del dispositivo, el cansancio del individuo entre los diferentes días... aun así existen estrategias como la técnica de fusión o la combinación de los distintos sensores, que parecen permitir obtener resultados razonablemente buenos en ambos dispositivos.

	DF-X (MICRO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.1148	0.2943	0.1148
2	0.1166	0.2402	0.1166
3	0	0.1625	0
5	0.0432	0.0253	0.0253
8	0.2922	0.0466	0.0466
9	0.1475	0.0916	0.0916
10	0.1445	0.3333	0.1445
11	0.0418	0.0544	0.0418
12	0.0779	0.3116	0.0779
13	0.1175	0.1705	0.1175
14	0.1351	0.2115	0.1351
15	0.3554	0.2706	0.2706
17	0.1088	0.1686	0.1088
18	0	0.1269	0
MEDIA	0.1211	0.1791	0.0922
	10	4	

(a) Pulsera Microsoft

	DF-X (MOTO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.0909	0.1016	0.0909
2	0.0536	0.0813	0.0536
3	0.0643	0.2593	0.0643
5	0.0417	0.0072	0.0072
8	0.2044	0.1733	0.1733
9	0.1246	0.1273	0.1246
10	0.4328	0.2967	0.2967
11	0.0167	0.1667	0.0167
12	0.077	0.1685	0.077
13	0.1568	0.2985	0.1568
14	0.1204	0.1049	0.1049
15	0.336	0.1061	0.1061
17	0.1818	0.2108	0.1818
18	0.0248	0.0415	0.0248
MEDIA	0.1375	0.1531	0.1056
	9	5	

(b) Reloj Motorola

Tabla 6.18: Combinación ACC/GYR de la componente X en el Dominio de la Frecuencia.

	DF-Y (MICRO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.3951	0.0682	0.0682
2	0.3817	0.2363	0.2363
3	0.15	0.1654	0.15
5	0.0943	0.2063	0.0943
8	0.1667	0.2883	0.1667
9	0.2072	0	0
10	0.1886	0.2523	0.1886
11	0.2163	0.141	0.141
12	0.2781	0.3	0.2781
13	0.1768	0.0816	0.0816
14	0.0741	0.1451	0.0741
15	0.4197	0.3423	0.3423
17	0	0	0
18	0.0769	0.2796	0.0769
MEDIA	0.2018	0.179	0.1356
	7.5	6.5	

(a) Pulsera Microsoft

	DF-Y (MOTO)		MEJOR
	ACC	GYR	
	MultiMulti	MultiMulti	
1	0.2849	0	0
2	0.2483	0.24	0.24
3	0.191	0.0902	0.0902
5	0.0923	0.0306	0.0306
8	0.1798	0.2413	0.1798
9	0.0546	0.0419	0.0419
10	0.2066	0.2072	0.2066
11	0.0938	0.2975	0.0938
12	0	0.2249	0
13	0.339	0.2829	0.2829
14	0.1795	0.0558	0.0558
15	0.3843	0.4234	0.3843
17	0.0645	0.2533	0.0645
18	0.0601	0.1156	0.0601
MEDIA	0.1699	0.1789	0.1236
	7	7	

(b) Reloj Motorola

Tabla 6.19: Combinación ACC/GYR de la componente Y en el Dominio de la Frecuencia.

DF-Z (MICRO)				DF-Z (MOTO)			
	ACC	GYR		ACC	GYR		
	MultiMulti	MultiMulti	MEJOR	MultiMulti	MultiMulti	MEJOR	
1	0.2429	0.3613	0.2429	0.1176	0.3353	0.1176	
2	0.0465	0.079	0.0465	0.1525	0.1889	0.1525	
3	0.3889	0.2489	0.2489	0.1028	0.3022	0.1028	
5	0.1239	0.1012	0.1012	0.0314	0.019	0.019	
8	0.1508	0.2848	0.1508	0.3565	0.2122	0.2122	
9	0.1386	0.2819	0.1386	0.0488	0.2107	0.0488	
10	0	0.2631	0	0.02	0.2708	0.02	
11	0.016	0.1761	0.016	0.0364	0.0711	0.0364	
12	0.2232	0.1822	0.1822	0.2619	0.2065	0.2065	
13	0.2447	0.1111	0.1111	0.3384	0.2205	0.2205	
14	0.1204	0.2953	0.1204	0.0636	0.2937	0.0636	
15	0.1382	0.4056	0.1382	0.3288	0.4335	0.3288	
17	0.0773	0.3252	0.0773	0.2288	0.1025	0.1025	
18	0.1528	0.2628	0.1528	0.0851	0.1381	0.0851	
MEDIA	0.1475	0.2413	0.1234	0.1552	0.2146	0.1226	
	10	4		9	5		

(a) Pulsera Microsoft

(b) Reloj Motorola

Tabla 6.20: Combinación ACC/GYR de la componente Z en el Dominio de la Frecuencia.

Componente	Usuarios con mayor EER MICRO	Usuarios con mayor EER MOTO
X	10, 14, 15	8, 9, 10, 13, 17
Y	2, 3, 8, 10, 11, 12, 15	2, 8, 10, 13, 15
Z	1, 3, 8, 9, 12, 14, 15, 18	2, 8, 12, 13, 15
módulo	8, 10, 14, 15	3, 8, 9, 10, 15

Tabla 6.21: Resumen Usuarios EER alto Combinación ACC/GYR - Dominio de la Frecuencia.

6.3. Comparación entre dominios

En el estudio detallado de cada uno de los dominios, por separado, se ha podido ver como la combinación de sensores beneficiaba en todos los casos, consiguiendo mejores resultados. Las tablas 6.22 y 6.23 muestran las tasas de equierror medias en el escenario estudiado de *Multisesión-Multimuestra* para ambos sensores y su combinación, viéndose de manera conjunta en ambos dominios, para poder estudiar las diferencias que existen entre los dominios, una vez conseguidas las mejoras. Cada fila representa cada una de las componentes estudiadas: X, Y, Z y módulo.

En la pulsera *Microsoft*, el módulo es lo que mejor funciona en ambos dominios, consiguiendo resultados bastante buenos en cada uno de los sensores, pero aún mejores tras su combinación, llegando a un valor medio de 6.04% en el Dominio del Tiempo y 8.87% en el Dominio de la Frecuencia. En el reloj *Motorola*, en el Dominio del Tiempo funcionan bien tanto las componentes individuales como el módulo con resultados bastante similares pero consiguiendo su mejor EER medio en el módulo, de un 7.61% tras la combinación de ambos sensores. En cambio, en el Dominio de la Frecuencia, los mejores resultados se consiguen con la componente X, tanto de los sensores individuales como tras la combinación, consiguiendo un EER medio de 0.1056%, mientras que utilizar lo mismo con el módulo implica empeorar la tasa de equierror media en algo menos que un 2%. Por otro lado, se puede ver que en ambos dispositivos funciona mejor el Dominio del Tiempo, con diferencias inferiores al 3% con el Dominio de la Frecuencia. También se puede decir que funciona mejor la pulsera de *Microsoft*, con diferencias inferiores al 2%. No obstante, el estudio se ha hecho con pocos usuarios y sería apropiado repetirlo para un Base de Datos más grande para obtener conclusiones más fiables.

Componente MICRO	ACC DT	GYR DT	ACC+GYR DT	ACC DF	GYR DF	ACC+GYR DF
X	0.1347	0.1483	0.0847	0.1211	0.1791	0.0922
Y	0.1804	0.1299	0.0959	0.2018	0.179	0.1356
Z	0.129	0.171	0.1198	0.1475	0.2413	0.1234
módulo	0.1241	0.1026	0.0604	0.1175	0.1681	0.0887

Tabla 6.22: Resumen de la combinación de sensores en el dispositivo MICRO.

Componente MOTO	ACC DT	GYR DT	ACC+GYR DT	ACC DF	GYR DF	ACC+GYR DF
X	0.1142	0.1801	0.0773	0.1375	0.1531	0.1056
Y	0.1335	0.124	0.0813	0.1699	0.1789	0.1236
Z	0.1336	0.1307	0.0805	0.1552	0.2146	0.1226
módulo	0.1217	0.1508	0.0761	0.1549	0.1857	0.1233

Tabla 6.23: Resumen de la combinación de sensores en el dispositivo MOTO.

En las tablas 6.11 y 6.21 se indicaban aquellos usuarios con EER alto. Se desconocen las razones por las que los usuarios malos cambian entre sensores y dispositivos, pero sí se puede apreciar una tendencia en la que los usuarios 3 y 9 se repiten dentro de la pulsera *Microsoft*, los usuarios 10 y 18 en el reloj *Motorola*, mientras que los usuarios 8 y 15 son aparentemente malos en ambos dispositivos. Lo que si es cierto y ocurre en todas las biometrías es que los sistemas fallan porque el comportamiento humano es variable. En el caso de la huella dactilar, si no sitúas bien el dedo o no hay suficiente luz en la habitación, el sistema podría no reconocerte, aquí, si no te colocas bien la pulsera, tienes prisa o estas más cansado de lo habitual, el sistema podría confundirte con otro usuario y, en consecuencia, no reconocerte consiguiendo un EER alto. Este problema es bien conocido en biometría desde 1998, cuando se publicó el primer artículo [73] que describía lo que actualmente se conoce como “animalario”. En él, categoriza a los seres humanos en 4 categorías.

- **Ovejas:** aquellos usuarios buenos, fáciles de reconocer.
- **Cabras:** usuarios difíciles de reconocer.
- **Corderos:** usuarios fáciles de imitar, siendo muy probable que una persona elegida al azar sea aceptada como cordero.
- **Lobos:** personas particularmente exitosas en imitar a otras personas, siendo muy probable que su patrón sea aceptado como el de otra persona.

Tanto las cabras como los corderos y los lobos afectan negativamente al rendimiento de los sistemas, y es que lo más probable es que la población de usuarios exhiba características continuas de estos 3 animales. El documento realiza pruebas estadísticas para la existencia de estos animales y lo aplica a la evaluación del reconocimiento de la forma de hablar. No obstante, existen numerosos artículos que muestran el mismo comportamiento en otras biometrías.

Capítulo 7

Adquisición de nuevos datos

La disponibilidad de únicamente 20 usuarios había planteado la posibilidad de recoger nuevos y más datos, pero en un principio se había descartado y decidido aprovechar la potencia de tener pocos usuarios para hacer un análisis más detallado de cada uno de ellos, que con muchos no habría sido abarcable.

El problema de que para 8 usuarios se tuvieran dos muestras de datos; mientras que para 11 usuarios se tenían cuatro y para 1 se tenían tres, por la necesidad de eliminar una toma de datos del usuario 13, ha hecho que en la forma experimental de Monosesión-Monomuestra, cuando se trabaja únicamente con la sesión 1 de los datos pero con la necesidad de que todos los usuarios utilizados tengan datos en la muestra 1 y en la muestra 2 de la sesión 1, se disponga de 14 usuarios que son bastantes pocos.

Además, la ausencia de información sobre cómo fueron tomados los datos y si de verdad los usuarios estaban andando normal o si hubo algunas alteraciones, que pudieran explicar por qué unos usuarios eran mejores que otros, ha llevado en la iteración 7 a poner en marcha la aplicación *Android* que permite recoger datos, a pesar de la previsión de tiempo que eso suponía y la posibilidad de tener que abandonarlo sin éxito.

7.1. Puesta en marcha de los dispositivos y la APK

Lo primero de todo, fue comprobar que los dos dispositivos a los que se tenía acceso seguían funcionando. Estos dispositivos son:

1. Pulsera Microsoft Band 2.
2. Reloj Motorola Moto 360 SPORT.

Tras cargarlos y verificarlo, se vinculó ambos dispositivos al teléfono móvil con el que se trabajaría: LG Q6 con versión *Android 8.1.0*. Al encender el reloj se realizaron una serie de actualizaciones que lo llevaron a la versión de Android 7.1.1.

Para vincular los dispositivos, puesto que el teléfono de destino era diferente al que se utilizó la última vez que se recogieron datos, era necesario restaurarlos a los valores de fábrica siguiendo los siguientes pasos:

1. En la pulsa de Microsoft: Ajustes/Símbolo de apagar/Restab. de fábrica/Restablecer disp.
2. En el reloj Motorola: Ajustes/Sistema/Desconectar y reestablecer.

Para ello, también es necesario instalar en el teléfono móvil las aplicaciones: Microsoft Band y Wear OS, disponibles en la *Play Store*. Activar el Bluetooth en los 3 dispositivos y seguir los pasos de las propias aplicaciones de cada dispositivo para vincularlos.

Después, la primera opción fue instalar en el teléfono móvil el archivo *WearableSensors.apk* que se tenía del anterior TFG [4], lo que llevo a que la pulsera de Microsoft sí que funcionase, recogiendo los datos correctamente al ponerlo en marcha pero el reloj Motorola, el cuál dependía de Android, no. Por tanto, se instaló *Android Studio* en el ordenador y se recuperaron los archivos fuentes con los que se había programado la aplicación.

Todo el rato se va a trabajar con un ordenador con Windows 10.

Se empezó abriendo el archivo *build.gradle*, que se encontraba dentro de la carpeta principal de los archivos fuente de la aplicación, en *Android Studio*. Una vez abierto, el proyecto está cargado y se tienen que tener las carpetas mobile y wear junto con los archivos gradle (Gradle Scripts) que forman el proyecto tal y como se ve a la izquierda de la figura 7.1.

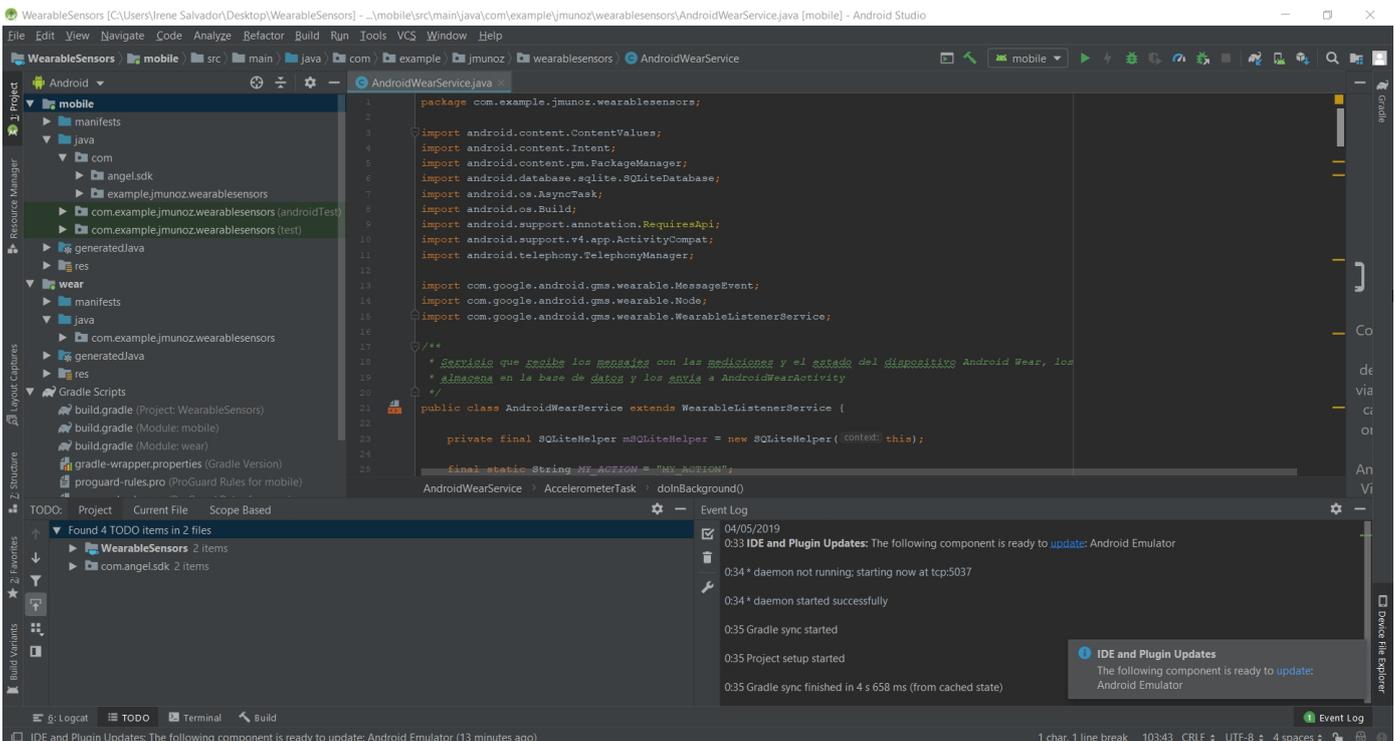


Figura 7.1: Visión Android Studio.

Evidentemente después de varios años, había muchos errores. Una manera de quitarlos y la que se ha seguido para conseguir que las cosas funcionasen consta de los siguientes pasos¹:

- Abrir los archivos *build.gradle*. Se tienen 3: uno perteneciente al proyecto (Project: Wearable-Sensors) y otros dos pertenecientes a cada uno de los módulos: (Module: mobile) y (Module: wear). Modificar en ellos cada una de las cosas que aparezcan subrayadas en rojo. Cada vez que hay algo rojo, se sitúa una bombilla roja al lado de la línea de código, proporcionando algunas de las opciones posibles para solucionar el problema. Empezar con los parámetros *compileSdkVersion* y *targetSdkVersion* (si aparecen en rojo). En este caso, te aconseja la versión que debes usar. Para este proyecto se ha utilizado la versión 28, mientras que en el anterior TFG [4] se usaba la 23. Después, intentar corregir cada línea en rojo de estos archivos con ayuda de la bombilla o buscando el error en Internet.
- Ir abriendo cada una de las clases del módulo *Mobile* dentro de la carpeta java e ir solucionando los diferentes errores subrayados en rojo de la misma manera que antes: con ayuda de la bombilla roja que proporciona *Android Studio* o buscando en Internet. Para saber si una clase tiene más errores, hay que mirar la parte superior derecha de la barra de desplazamiento de la clase, donde se puede ver, por ejemplo, un cuadrado amarillo como ocurre en la figura 7.1. Pueden aparecer las siguientes cosas:
 - Tic verde: no hay más errores.
 - Cuadrado amarillo: no hay más errores. Existen warnings.
 - Exclamación en un círculo rojo: sigue habiendo errores.
- Ir abriendo cada una de las clases del módulo *Wear* dentro de la carpeta java y actuar de la misma manera que en el paso anterior.

La mayoría de los errores de las clases serán debidos a cosas que han desaparecido en la versión de *Android* que estás utilizando y que en las anteriores sí existía. En algunos casos, se solucionará con una frase similar a: *@RequiresApi(api = api_correspondiente)* que te proporcionará la bombilla roja de ayuda en *Android Studio*.

Solucionados los errores, se ha construido el proyecto de la siguiente manera:

- Build>Clean Project
- Build>Make Project

Una vez construido con éxito y conseguido los mensajes: *Gradle sync started* y *Gradle sync finished in 1 min 4 s* que se pueden ver en la opción *TODO* que se muestra en la parte inferior de la figura 7.1, hay que hacer lo siguiente:

¹Puede que estos pasos no sean los mejores. Los conocimientos de *Android* de la persona encargada eran limitados, ya que era la primera vez que veía un proyecto de este estilo. Un experto en estas tecnologías probablemente actuase de una manera diferente y mejor.

- Conectar el teléfono móvil al ordenador utilizando la opción de conexión USB “Cargando” como se muestra en la figura 7.2.
- En el teléfono móvil ir a Ajustes/Sistema, permitir las opciones de desarrollador presionando 7 veces el número de compilación en la información del Software [74].
- En el teléfono móvil ir a las opciones de desarrollador, activar la opción “Depuración USB”.

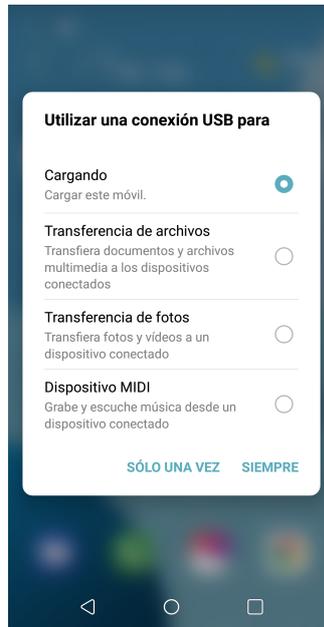


Figura 7.2: Opción USB conexión teléfono móvil-PC.

Ahora mismo, si en *Android Studio* ejecutamos el proyecto de la opción Run>Run, veremos nuestro teléfono móvil vinculado. No obstante, el reloj Motorola requiere tener instalada la aplicación Android y para ello, hay que seguir los siguientes pasos:

- En la aplicación Wear OS, en los “Ajustes avanzados”, activar la opción: “Depuración por Bluetooth”. Aparecerá algo como lo de la figura 7.3 (a).
- Aunque el destino está conectado, el host está desconectado y es necesario que se conecte. Para ello hay que buscar el archivo “adb.exe” de Android. En mi caso se encontraba en la ruta: *C:/Users/Irene Salvador/AppData/Local/Android/Sdk/platform-tools*. De no existir, sería necesario instalarlo [75]. Una vez situado en esa carpeta, abrir la consola de *Windows* y ejecutar lo siguiente [76, 77]:
 - adb.exe
 - adb forward tcp:4444 localabstract:/adb-hub
 - adb connect localhost:4444
- Tras ejecutar las ordenes en la consola, el host aparecerá como conectado, como en la figura 7.3 (b).

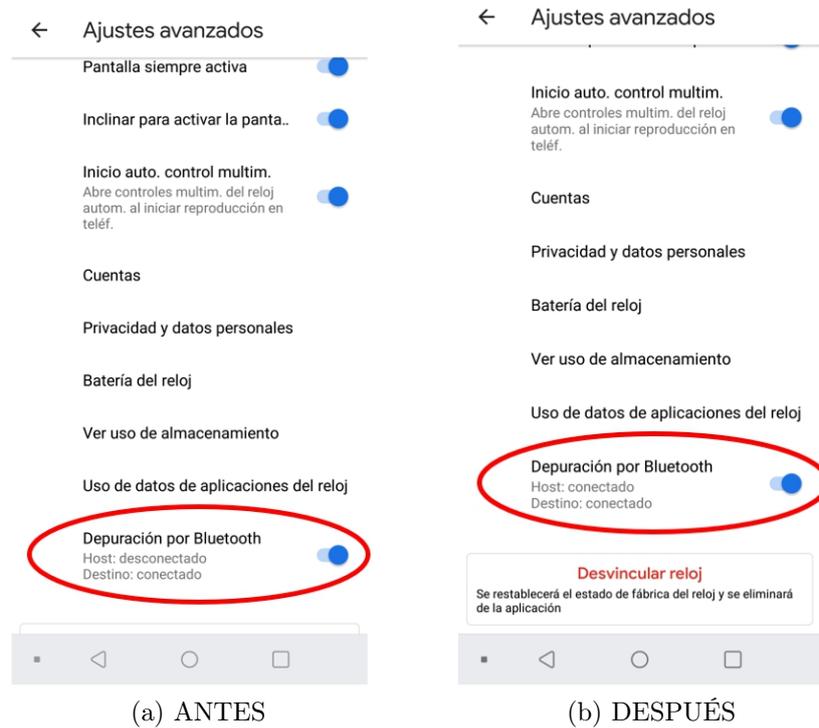


Figura 7.3: Depuración por Bluetooth Móvil-Reloj.

Ahora sí, en Android Studio aparecerá tanto el teléfono móvil como el reloj Motorola. Ejecutar la opción “WearActivity” en el reloj haciendo *Run>Run>WearActivity>Motorola Moto 360* y la opción “mobile” en el teléfono móvil haciendo: *Run>Run>mobile>LGE LG-M700*. Al ejecutarlo en el reloj se construirá la aplicación a la que se puede acceder desde Ajustes/Aplicaciones/WearableSensors. En el móvil también se creará, abriéndose directamente, como puede verse en la figura 7.4.

Ahora sólo hay que elegir la opción en la aplicación del teléfono móvil: *Microsoft Band* o *androidwear*, introducir los datos: usuario, sesión, tarea y muestra, presionar el botón “ACEPTAR”, después “EMPEZAR” y comenzarán a recogerse datos como puede verse en la figura 7.5 a y b.

Para recoger datos, no es necesario tener el móvil conectado al ordenador, ni las opciones de depuración activadas. Ya está instalada la aplicación y una vez funciona, no es necesario utilizar Android Studio para nada más.

Al final, por falta de tiempo, y a pesar del invertido en poner en marcha la aplicación, no se han realizado nuevas capturas de datos, ya que implicaba encontrar a las suficientes personas y reunirlos 2 días diferentes. Se va a dejar como trabajo futuro, que no será complicado, ya que está en marcha tanto la aplicación Android como los credenciales necesarios para acceder a la Base de Datos.

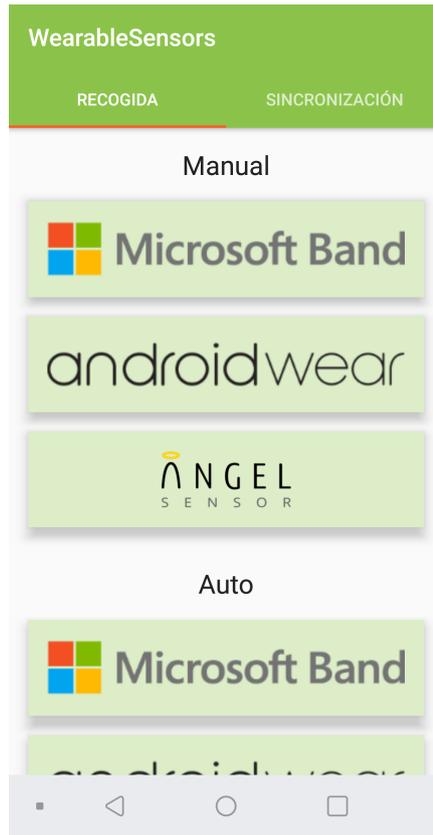
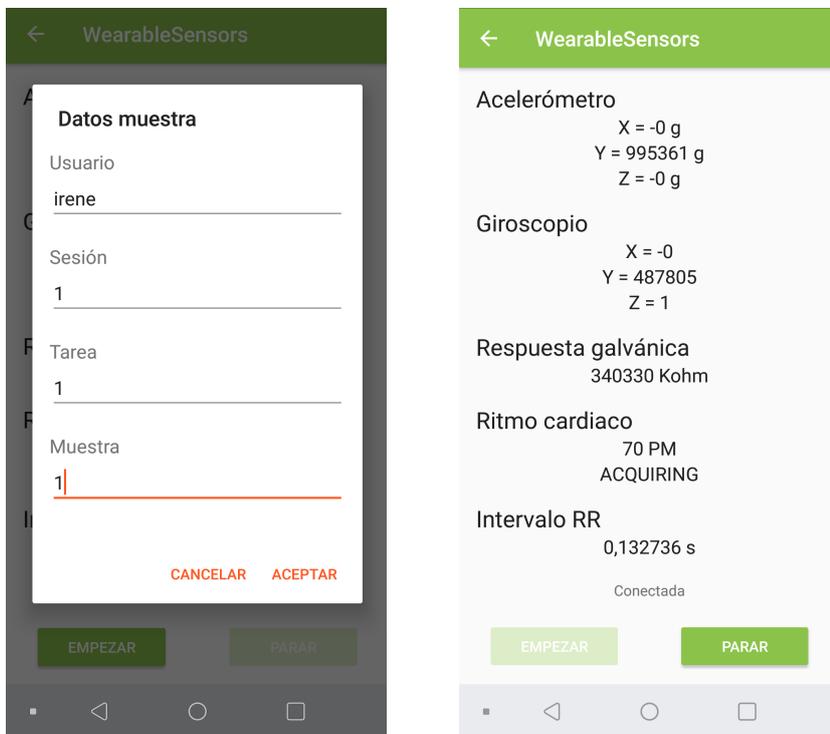


Figura 7.4: Menú aplicación WearableSensors desde el teléfono móvil.



(a) ANTES

(b) DESPUÉS

Figura 7.5: Aplicación recogiendo datos

Capítulo 8

Conclusiones y trabajo futuro

8.1. Conclusiones

El trabajo expuesto permite concluir que se han cumplido todos los objetivos inicialmente planteados.

Se ha realizado un análisis visual de la señal original, el cual ha permitido observar anomalías y corregirlas, optando por su eliminación manual.

Con la señal limpia, se han aprovechado los análisis realizados en [5], para extraer el sistema de reconocimiento final construido con las opciones de preprocesamiento y el análisis de los parámetros, que conseguía optimizar y reducir el problema, permitiendo la realización de una aproximación inicial. Se ha probado en ambos sensores, acelerómetro y giroscopio, evaluando sus diferencias y demostrando ser similares y cómo una combinación de ambos podría generar mejores resultados.

También se han contrastado los resultados con respecto al tipo de ponible usado evaluándolo en los diferentes escenarios creados en el diseño experimental (*Monosesión-Monomuestra*, *Multisesión-Monomuestra*, *Multisesión-Multimuestra*). Las diferencias, considerando el entorno experimental más realista, *Multisesión-Multimuestra* y el módulo de los datos, no han superado el 2% en ninguno de los dos dominios y sensores. Esto muestra la posible independencia del sistema de reconocimiento con respecto al dispositivo.

Se puede concluir que los resultados obtenidos son prometedores y muestran que el uso de dispositivos ponibles puede ser una alternativa muy interesante, aunque con muchas cuestiones todavía abiertas. Para poder abordar este estudio es imprescindible tener una base de datos más completa. Aunque no ha dado tiempo a ello, sí que se ha preparado el entorno de adquisición, dejándose en funcionamiento para poder realizar esa nueva adquisición de datos y continuar con el trabajo.

En cuanto a las conclusiones personales, este trabajo me ha permitido poner en práctica parte de los conocimientos adquiridos en diversas asignaturas a lo largo de mis estudios, aprender

nuevos conceptos, el diseño experimental y la manera de trabajar en biometría, el conocimiento y funcionamiento de problemas de clasificación trabajando con muestras de la misma clase o *one-class*, ampliar y reforzar mis habilidades con la herramienta de programación R, así como vivir la experiencia de un problema actual y novedoso que no tiene una solución única, sino una infinidad de posibilidades con las que poder seguir trabajando.

Por otro lado, este trabajo me ha ayudado a organizarme mejor, sacando cada semana las cosas que quedaban por hacer y realizándolas de manera progresiva. Me ha permitido trabajar en un grupo de investigación y entender mejor cómo funciona el proceso de la investigación científica desde dentro. Habiendo conseguido, con todo ello, una experiencia muy productiva y enriquecedora, tanto de manera personal como académica.

8.2. Líneas de trabajo futuro

En este proyecto se ha visto la aplicación del sistema de reconocimiento construido en [5], permitiendo ambos proyectos sentar unas bases sólidas sobre las que poder empezar a trabajar en esta biometría, pero aún existe bastante trabajo por delante. Los resultados obtenidos mediante nuestro sistema de reconocimiento final construido han sido buenos y prometedores, pero se ha utilizado una Base de Datos pequeña. Ahora, el primer paso a abordar es recoger más datos, necesitándose al menos 25 usuarios diferentes. El proceso para su recolección tiene que ser muy estricto:

- Capturar datos durante al menos 2 sesiones (días diferentes). Si fuese posible, sería muy interesante capturar datos en una tercera sesión, pasado bastante tiempo, por ejemplo, 2 meses de las anteriores.
- En cada sesión, recoger dos muestras de datos a cada usuario. Cada muestra consiste en realizar el recorrido dos veces, uno con la pulsera de *Microsoft* y otra con el reloj *Motorola*. De nuevo, si fuera posible, sería interesante capturar los datos dos veces, una portando el dispositivo en la mano dominante, y otra en la contraria.
- Realizar un recorrido más largo al actualmente utilizado de, por ejemplo, 5 minutos. Incluso si fuese posible, realizar una nueva sesión en la que el usuario lleve el dispositivo a lo largo de una mañana entera.

Por otro lado, como se vio en el estado del arte, existen dos maneras de trabajar, extrayendo características sobre los marcos temporales en que se divide la muestra o utilizando la señal preprocesada cruda. Aquí, basándonos en la bibliografía y en pruebas simples realizadas a los datos utilizando el método DTW a la señal cruda, se decidió extraer características, por conseguir, en el dominio del tiempo resultados bastante mejores. No obstante, es una vía sobre la que poder trabajar, utilizando, en el dominio de la frecuencia, directamente las amplitudes del Análisis de Fourier en lugar de sus características.

Idealmente, como trabajo futuro, se tendría que buscar una manera de optimizar el sistema a las características de cada individuo, ya que como se ha visto a lo largo de esta memoria y [5], si fuésemos capaces de predecir la componente tridimensional mejor de cada individuo, los resultados mejorarían mucho. Incluso, simplemente siendo capaces de predecir el tipo de sensor que va a funcionar mejor, aunque por esta vía se deja pendiente la prueba de una combinación de ambos sensores.

Pero sin duda, a la vista de los resultados obtenidos, el presente TFG planta una semilla muy interesante para futuros trabajos, bajo la necesidad de capturar más datos.

Acrónimos y abreviaturas

ACC Acelerómetro	I identificación
AI Inteligencia Artificial	ICA Independent Component Analysis
AI HLEG Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial	ID Identificador
DCT Transformada del Coseno Discreta	IMEI International Mobile Equipment Identity
DF Dominio de la Frecuencia	KNN K-Nearest Neighbors
DT Dominio del Tiempo	LDA Linear Discriminant Analysis
DTree Decision Trees	MICRO Microsoft
DTW Dynamic Time Warping	ML Machine Learning
ECTS European Credit Transfer and Accumulation System	MOTO Motorola
EER tasa de equierror	MonoMono Monosesión-Monomuestra
EU Unión Europea	MultiMono Multisesión-Monomuestra
FNR tasa de falsos negativos	MultiMulti Multisesión-Multimuestra
FPR tasa de falsos positivos	PCA Principal Component Analysis
GMM Gaussian Mixture Model	PPG sensores fotopleletismográficos
GPS Sistema de Posicionamiento Global	ROC Receiver Operating Characteristic
GYR Giroscopio	SVM Support Vector Machines
H precisión	TFG Trabajo Fin de Grado
HMM Hidden Markov Model	UVa Universidad de Valladolid
	V verificación

Índice alfabético

- Análisis de Fourier, 42, 43, 70, 80, 96, 122
- Base de Datos, 19, 27, 28, 30, 32, 35, 40, 47, 58, 61, 63–66, 83, 112, 119, 122
- Biometría, 5, 18, 19, 21–24, 27, 45, 48, 55, 56, 58, 59, 76, 82, 84–86, 102, 104, 109, 113, 122
- Dispositivos comerciales, 5, 18, 23, 45, 56, 61, 64, 71
- Dispositivos ponibles, 5, 18, 23, 31, 42, 55, 56, 61, 76
- Dominio de la Frecuencia, 30–34, 42, 43, 48, 49, 57, 58, 76, 78, 89–91, 95, 96, 104–106, 108–112, 122
- Dominio del Tiempo, 30–34, 42, 43, 48, 49, 57, 58, 76, 80, 89, 90, 95–99, 101–105, 109, 112, 122
- Dygraphs, 42, 53
- Forma de andar, 18, 20, 21, 23, 42, 56, 57
- INdat, 30
- Iteración, 25, 26, 28, 37, 40, 42–47, 49, 50
- PMBOK, 27
- Reconocimiento biométrico, 18, 20, 23, 55, 58, 76
- SCRUM, 25, 26, 28, 29
- Selección de características, 30, 31, 43, 44, 48, 50, 76
- UVa, 27, 51

Anexos

Apéndice A

Contenido del CD

/	
— memoria.pdf	Memoria del Trabajo de Fin de Grado en Ingeniería Informática
— Programas	.
— construyeTablasCaracteristicas.R.....	Script utilizado para construir las tablas de características.
— RendimientoFinal_KNN_tamanoFijo.R.....	Script que construye las tablas completas del rendimiento final considerando un tamaño fijo de ventana.
— RendimientoFinal_KNN_secuenciado.R.....	Script que construye las tablas completas del rendimiento final considerando un rango de tamaños de ventana.
— Imágenes	Carpeta con todas las imágenes utilizadas en este documento. Por cada capítulo, se tiene una carpeta con su nombre y las imágenes correspondientes.

Bibliografía

- [1] Alimarket Alimentación, 28 de enero de 2019, “Minsait lanza una solución para la compra online de frescos”, <https://www.alimarket.es/alimentacion/noticia/292022/minsait-lanza-una-solucion-para-la-compra-online-de-frescos>
- [2] Donweb agencia dde prensa, 2014, “Los 14 usos de drones que seguro no conocías”, <http://agencia.donweb.com/los-14-usos-de-drones-que-seguro-no-conocias/>
- [3] J. M. Galán, TFG de la Escuela de Ingeniería Informática de la Universidad de Valladolid curso 2015/2016, “Wearables: Análisis de dispositivos y recogida de datos en Android para estudios biométricos”
- [4] Daniel González Alonso, TFG de la Escuela de Ingeniería Informática de la Universidad de Valladolid curso 2016/2017, “Estudio preliminar del uso de Wearables en reconocimiento biométrico de personas”
- [5] Irene Salvador Ortega, 26 de junio de 2019, “Extracción de características y clasificación para la implementación de un sistema de reconocimiento biométrico mediante dispositivos ponibles (wearables)”
- [6] Jorge Blasco, septiembre de 2016, “A Survey of Wearable Biometric Recognition Systems”
- [7] Anil K. Jain, 2004, “Multibiometric systems. Communications”, ACM
- [8] Roman V. Yampolskiy, 2010, “Taxonomy of behavioral biometrics”, Behavioral Biometrics for Human Identification
- [9] Salil Prabhakar, 2003, “Biometric recognition: Security and privacy concerns”, IEEE Security & Privacy
- [10] European Commission’s High-Level Expert Group on Artificial Intelligence, 18 de diciembre de 2018, “Draft Ethics guidelines for trustworthy AI”, <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>
- [11] Ten Kettles Inc., 2014-2019, página oficial, <http://www.tenkettles.com/>

- [12] Alex Andrews, 2 de junio de 2017, “Scrum Of One: How to Bring Scrum into your One-Person Operation”, <https://www.raywenderlich.com/585-scrum-of-one-how-to-bring-scrum-into-your-one-person-operation>
- [13] HMD Project Managers, 22 de agosto de 2017, “Cuales son los Principios básicos de SCRUM”, <https://uv-mdap.com/blog/principios-basicos-de-scrum-metodologias-agiles/>
- [14] Adri Salazar, 2 de octubre de 2016, “Principios de Scrum”, <http://www.prozessgroup.com/principios-de-scrum/>
- [15] Pablo de la Fuente Redondo, diapositivas para el curso 2018/2019, asignatura *Planificación y Diseño de Sistemas Computacionales*, Universidad de Valladolid
- [16] Project Management Institute, 2008, “A guide to the project management body of knowledge (PMBOK Guide)”, Ed. Newton Square. Enlace web: <https://www.pmi.org/pmbok-guide-standards/foundational/pmbok>
- [17] Sergio, 2019, “Plantillas calendario en Excel 2019”, <https://ayudaexcel.com/plantillas-calendario-en-excel-2019/>
- [18] Daniel González Alonso, última visualización el 30 de enero de 2019, “visualizacion-wearables”, <https://github.com/Zalez95/visualizacion-wearables/>
- [19] (s.a), extraído el 20 de febrero de 2019, “dygraphs for R”, <https://rstudio.github.io/dygraphs/> (<https://github.com/rstudio/dygraphs/issues/29>)
- [20] LifeSizeCloud, plataforma de videoconferencias, último acceso el 20 de junio de 2019, <https://call.lifesizecloud.com/>
- [21] WeTransfer, almacenamiento de documentos, último acceso el 20 de junio de 2019, <https://wetransfer.com/>
- [22] PhotoJoiner, edición de imágenes, último acceso el 20 de junio de 2019, <https://old.photojoiner.net/>
- [23] Fotor, edición de imágenes, último acceso el 20 de junio de 2019, <https://www.fotor.com/>
- [24] Rstudio, entorno de desarrollo, último acceso el 20 de junio de 2019, <https://www.rstudio.com/>
- [25] Anaconda, entorno de desarrollo, último acceso el 20 de junio de 2019, <https://www.anaconda.com/distribution/>
- [26] PandasPython, biblioteca de Python, último acceso el 20 de junio de 2019, <https://www.learnpython.org/es/Pandas%20Basics>

- [27] SklearnPython, biblioteca de Python, último acceso el 20 de junio de 2019, <https://scikit-learn.org/stable/>
- [28] Matplotlib, biblioteca de Python, último acceso el 20 de junio de 2019, <https://matplotlib.org/>
- [29] Joe Belfiore, 2015, “Making Windows 10 More Personal and More Secure with Windows Hello”, <https://blogs.windows.com/windowsexperience/2015/03/17/making-windows-10-more-personal-and-more-secure-with-windows-hello/>
- [30] Toshiyo Tamura, 2014, “Wearable photoplethysmographic sensors—Past and present”, *Electronics*.
- [31] Davide Maltoni, 2009, “Handbook of Fingerprint Recognition”, Springer.
- [32] Lucas Introna and Helen Nissenbaum, 2010, “Facial recognition technology a survey of policy and implementation issues. Technical Report”, The Department of Organisation, Work and Technology, Lancaster University.
- [33] Pallavi Meharia and Dharma P. Agrawal, 2015, “Unobtrusive gait verification for mobile phones”, *Journal of Information Privacy & Security*
- [34] Hong Lu, 2014, “Unobtrusive gait verification for mobile phones”, ACM
- [35] Mohammad Derawi, 2015, “Wireless chest-based ECG biometrics”, Springer
- [36] Hindra Kurniawan, 2013, “Stress detection from speech and galvanic skin response signals”, *IEEE 26th International Symposium on Computer-Based Medical Systems*
- [37] Adam S. Venable, 2013, “Gender differences in skin and core body temperature during exercise in a hot, humid environment”, *Internal Journal of Exercise Science: Conference Proceedings*, Vol. 2. 9.
- [38] Kenneth Revett, 2008, “Behavioral Biometrics: A Remote Access Approach”
- [39] Hoang Minh Thang, 2012, “Gait Identification Using Accelerometer on Mobile Phone”
- [40] Mohammad Omar Derawi, 2010, “Accelerometer-Based Gait Analysis, A survey”
- [41] Oresti Banos, 9 de abril de 2014, “Window Size Impact in Human Activity Recognition”
- [42] Guannan Wu, 10 de junio de 2018, “A Continuous Identity Authentication Scheme Based on Physiological and Behavioral Characteristics”
- [43] Akram Bayat, 2017, “Classifying Human Walking Patterns using Accelerometer Data from Smartphone”
- [44] Liu Yiyan, noviembre de 2016, “An Hidden Markov Model based Complex Walking Pattern Recognition Algorithm”

- [45] Weitao Xu, 2017, “Gait-Watch: A Context-aware Authentication System for Smart Watch Based on Gait Recognition”
- [46] Samer K Al Kork, 2017, “Biometric Database for Human Gait Recognition using Wearable Sensors and a Smartphone”
- [47] Sherif Said, 26 de julio de 2018, “Experimental Investigation of Human Gait Recognition Database using Wearable Sensors”
- [48] Davrondzhon Gafurov and Einar Snekkenes, 26 de abril de 2009, “Gait Recognition Using Wearable Motion Recording Sensors”
- [49] Thomas Bernecker, (s.f), “Activity Recognition on 3D Accelerometer Data (Technical Report)”
- [50] Fan Yang, 2017, “Real-Time Human Activity Classification by Accelerometer Embedded Wearable Devices”
- [51] Liu Rong, 2017, “A Wearable Acceleration Sensor System for Gait Recognition”
- [52] Davrondzhon Gafurov, noviembre de 2006, “Biometric Gait Authentication Using Accelerometer Sensor”
- [53] Bing Sun, 2014, “Gait Characteristic Analysis and Identification Based on the iPhone’s Accelerometer and Gyrometer”
- [54] Miikka Ermes, 2006, “Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions”
- [55] Chiung Ching Ho, 2010, “An Unobtrusive Android Person Verification Using Accelerometer Based Gait I”
- [56] Martin Reese Hestbek, 13 de abril de 2012, “Biometric Gait Recognition For Mobile Devices Using Wavelet Transform And Support Vector Machines”, 19th International Conference on Systems, Signals and Image Processing
- [57] Michael Fitzgerald Nowlan, 2009, “Human Identification via Gait Recognition Using Accelerometer Gyro Forces”
- [58] Heikki Ailisto, 2005, “Identifying people from gait pattern with accelerometers”
- [59] Pierluigi Casale, 2012, “Personalization and user verification in wearable systems using biometric walking patterns”, Personal and Ubiquitous Computing.
- [60] Ghina Dandachi, 2013, “A novel identification/verification model using smartphone’s sensors and user behavior”, 2nd International Conference on Advances in Biomedical Engineering

- [61] Chiung Ching Ho, 2012, “An unobtrusive Android person verification using accelerometer based gait II”, 10th International Conference on Advances in Mobile Computing & Multimedia
- [62] RashaWahid, 2012, “A Gaussian mixture models approach to human heart signal verification using different feature extraction algorithms”, Computer Applications for Bio-technology, Multimedia, and Ubiquitous City
- [63] Zhidong Zhao and Qinqin Shen, 2011, “A human identification system based on heart sounds and Gaussian mixture models”, 4th International Conference on Biomedical Engineering and Informatics
- [64] Zoubin Ghahramani, 2001, “An introduction to hidden Markov models and Bayesian networks”, International Journal of Pattern Recognition and Artificial Intelligence
- [65] J. Ross Quinlan, 2014, “C4.5: Programs for Machine Learning”, Elsevier
- [66] Daisuke Sugimori, 2011, “A study about identification of pedestrian by using 3-axis accelerometer”.
- [67] Claudia Nickel, 2012, “Authentication of smartphone users based on the way they walk using k-NN algorithm”, 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing
- [68] Claudia Nickel, 2011, “Using hidden Markov models for accelerometer-based biometric gait recognition”, IEEE 7th International Colloquium on Signal Processing and Its Applications
- [69] Alexander J. Casson, 6 de diciembre de 2016, “Gyroscope vs. accelerometer measurements of motion from wrist PPG during physical exercise”
- [70] Wikipedia, última edición el 20 de febrero de 2019, “Curtosis”, <https://es.wikipedia.org/wiki/Curtosis>
- [71] Wikipedia, última edición el 24 de mayo de 2018, “Asimetría estadística”, https://es.wikipedia.org/wiki/Asimetr%C3%ADa_estad%C3%ADstica
- [72] João Neto, marzo de 2013, “Fourier Transform: A R Tutorial”, <http://www.di.fc.ul.pt/~jpn/-r/fourier/fourier.html>
- [73] George Doddington, 1998, “SHEEP, GOATS, LAMBS and WOLVES. A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation”
- [74] Android Studio Developers, último acceso el 20 de junio de 2019, “Configurar opciones para programadores en el dispositivo”, <https://developer.android.com/studio/debug/dev-options?hl=es-419>
- [75] shimp208, 2 de septiembre de 2018, “[TOOL]Minimal ADB and Fastboot”, <https://forum.xda-developers.com/showthread.php?t=2317790>

- [76] simoneluconi, 11 de febrero de 2015, “Install and Uninstall APK on your Moto 360 (Via Bluetooth)”, <https://forum.xda-developers.com/smartwatch/moto-360/guide-install-apk-moto-360-t3028067>
- [77] Android Studio Developers, último acceso el 20 de junio de 2019, “Debugging over Bluetooth”, <http://www.androiddocs.com/training/wearables/apps/bt-debugging.html>