UNIVERSIDAD CARLOS III DE MADRID

TRABAJO FIN DE GRADO

Detección de estrés en señales de voz

Autor:
ALBA MÍNGUEZ SÁNCHEZ

Supervisor:
CARMEN PELÁEZ MORENO

Grado en Ingeniería en Tecnologías de Telecomunicación

Departamento de Teoría de la Señal y Comunicaciones

20 de junio de 2017

«*Nadie podrá decir que fue suerte.*»

Road Ramos

Universidad Carlos III de Madrid

Resumen

Escuela Politécnica Superior Leganés
Departamento de Teoría de la Señal y Comunicaciones

Grado en Ingeniería en Tecnologías de Telecomunicación

Detección de estrés en señales de voz

by Alba MÍNGUEZ SÁNCHEZ

El estrés se ha convertido en uno de los factores más importantes para aquellas profesiones en las que la toma rápida de decisiones bajo situaciones de presión es la tarea principal. Igualmente, el estrés es el causante de una de las fobias más comunes entre los adultos: la glosofobia o el miedo a hablar en público.

La necesidad de controlar estas situaciones de tensión ha desembocado en el estudio de reconocimiento de emociones y estrés. Sin embargo, en los últimos años, las investigaciones para detectar estrés a partir de la voz se han realizado, de forma general, en entornos de laboratorio de los que se extraen conclusiones limitadas a la hora de analizar eventos de estrés reales.

En nuestro proyecto, utilizaremos una base de datos de estudiantes que realizan discursos en público mientras se les toman medidas sobre su ritmo cardíaco. Realizaremos la extracción de un set básico de características de la voz y generaremos etiquetas basadas estas medidas biométricas con el fin de realizar una detección entre estrés y no estrés lo más precisa posible.

Con el fin de conseguir esta precisión, se realizará una extracción de características a diferentes niveles de análisis; las etiquetas se basarán en distintos umbrales de decisión y, por último, se diseñarán diversos clasificadores, dando lugar a un porcentaje de acierto, en cualquiera de las combinaciones, superior al 80 %.

Abstract

Emotional stress has become one of the most important factors for those jobs where the main task is making quick decisions under pressure. In addition, stress is the cause of one of the most common phobias among adults: glossophobia or the fear of public speaking.

The need to control these situations of tension has developed the study of emotions and stress recognition. However, in recent years, researches about stress detection from speech has been done, in general, in laboratory environments. These improvements are limited when analyzing real stress events.

In our project, we will use a database of students who make a speech in public while measurements of their heart rate are taken. We will perform the extraction of a basic set of features from speech signals and generate labels based on these biometric measurements in order to perform a detection between stress and non-stress as accurate as possible.

In order to achieve this precision, the feature extraction will be performed at different analysis levels; labels will be based on different decision thresholds and, finally, different classifiers will be designed, giving a score, for any of the combinations, greater than 80 %.

Agradecimientos

En primer lugar, gracias a mi tutora, Carmen, por confiar en mí para realizar este proyecto. Por todo lo que he aprendido gracias a ella y de ella, por guiarme en todo momento y por hacerme creer en mí siempre que yo dejaba de hacerlo.

A papá y mamá, gracias por hacer posible, una vez más, que haya subido a una nueva plataforma. A mis hermanos, Ainhoa y Carlos, por apoyarme siempre que lo he necesitado y por aguantarme en los momentos más duros de este año.

A mis amigas, Dámaris, Marta, Miriam y Julia, por mostrar interés por lo que estaba haciendo, aunque no lo entendiéseis del todo, y hacerme sentir que valía la pena. Por hacerme olvidar, aunque fuera por momentos, el agobio y el estrés, nunca mejor dicho, con esos ratitos que son para nosotras.

Y a ti, Jesús, por ayudarme siempre con todo, incluso cuando te pido que dejes de hacerlo. Por aguantar mis quejas desde el primer día, por animarme cuando más lo necesitaba y escucharme siempre. Por enseñarme que todo esfuerzo tiene su recompensa, y por ser, en definitiva, el mejor modelo a seguir que puedo tener.

Índice general

Índice de figuras

Índice de cuadros

Índice de abreviaturas

VSA	**V**oice **S**tress **A**nalysis
SR	**S**peech **R**ecognition
ASR	**A**utomatic **S**peech **R**ecognition
STT	**S**peech **T**o **T**ext
TTS	**T**ext **T**o **S**peech
PS	**P**ublic **S**peaking
HR	**H**eart **R**ate
UTC	**U**niversal **T**ime **C**oordinated
STAI	**State- T**rait **A**nxiety **I**nventory
ECG	**E**lectro**c**ardio**g**ram
GSR	**G**alvanic **S**kin **R**esponse
EDR	**E**lectro**d**ermal **R**esponse
SCR	**S**kin **C**onductance **R**esponse
EMG	**E**lectro**m**yo**g**raphy
MFCC	**M**el **F**requency **C**epstral **C**oefficients
LPC	**L**inear **P**redictive **C**oding
FFT	**F**ast **F**ourier **T**ransform
TEO	**T**eager **E**nergy **O**perator
HMM	**H**idden **M**arkov **M**odel
MLP	**M**ulti-**l**ayer **P**erceptron
ANN	**A**rtificial **N**eural **N**etwork
SVM	**S**upport **V**ector **M**achine
RBF	**R**adial **B**asis **F**unction
bpm	**b**eats **p**er **m**inute
VAD	**V**oice **A**ctivity **D**etector

Capítulo 1

Introducción

El objetivo de este primer capítulo es introducir al lector al proyecto que se desarrollará en la memoria, ofreciéndole una visión global del contexto en el que se encuentra el mismo. Contiene, además, los principales objetivos y motivaciones que han hecho que este estudio se lleve a cabo, además de presentar el entorno socio-económico y el marco regulador en el que este encaja. Por último, se detalla la estructura que seguirá el contenido de la memoria para intentar facilitar todo lo posible su lectura.

1.1. Introducción

Las emociones constituyen una parte fundamental del ser humano debido a su gran influencia en la percepción y en las tareas cotidianas, como, por ejemplo, el aprendizaje, la comunicación e incluso la toma racional de decisiones.

Partiendo de esta base y combinándola con los numerosos avances tecnológicos de los últimos años, nace la computación afectiva *(Affective Computing)*, es decir, el estudio y desarrollo de sistemas y dispositivos que pueden reconocer, interpretar, procesar y simular los sentimientos humanos [1]. Esta ciencia se divide esencialmente en dos tecnologías: análisis de emociones en expresiones faciales y en voz. En relación con esta última, se ha comprobado empíricamente que los cambios en el sistema nervioso autónomo pueden alterar indirectamente el habla de una persona. En consecuencia, las tecnologías afectivas aprovechan esta información para reconocer la emoción del hablante.

Durante los últimos años se han realizado multitud de estudios sobre el reconocimiento automático de emociones que han permitido, de algún modo, parametrizarlas y saber qué diferencia unas de las otras. Sin embargo, para el caso de la detección de estrés, debido a la amplitud de su alcance y a que en la mayoría de casos se considera producto de una combinación de emociones, aún no se han hecho estudios suficientes para conocer qué características, medidas biométricas o psicológicas pueden ser más reveladoras para su análisis.

El estudio de la detección de estrés se puede emplear en numerosas aplicaciones. Un ejemplo de estas es el caso de la *glosofobia*, es decir, el miedo o incluso pánico a hablar en público, como en entrevistas de trabajo o discursos. Además de esto, también se puede aplicar a puestos de trabajo sometidos a situaciones de estrés en los que la rapidez en la toma de decisiones se convierte

en un requisito esencial. Este es el caso de, por ejemplo, los controladores aéreos, pilotos o el personal que gestiona el servicio de emergencias. Otro caso de detección podría darse incluso en situaciones de violencia doméstica, donde se buscaría la manifestación de estrés de la persona en peligro. En estas situaciones en las que aparece un estrés cognitivo tan importante, resulta necesario detectar cuándo una persona está sometida a ciertos niveles de tensión con el fin de llevar a cabo las acciones que correspondan.

A pesar de esta necesidad, las investigaciones que se han desarrollado hasta la fecha en este ámbito se han realizado bajo condiciones de laboratorio. En estos estudios, además de las grabaciones de voz, se recogen datos fisiológicos con el fin de ayudar a sacar conclusiones de las observaciones. Sin embargo, aunque los resultados obtenidos hasta la fecha han ayudado a ampliar el conocimiento en este área, proporcionan una generalización limitada a la hora de evaluar situaciones reales de estrés. Para superar esta limitación, se ha recomendado que las observaciones se realicen en situaciones de estrés objetivas, es decir, en eventos reales [2].

En nuestro estudio en concreto, abordamos la detección de estrés para un conjunto de estudiantes que realizan discursos en público de los cuales contamos con distintas grabaciones de voz y datos fisiológicos. La idea de este proyecto será enseñar a una máquina a realizar esta detección de estrés en señales de voz. Para ello, basaremos la anotación en valores de ritmo cardíaco y extraeremos de la voz un conjunto de características básicas, junto a sus estadísticos, con el fin de lograr la detección (clasificación binaria) entre estrés y no estrés.

1.2. Motivación y objetivos

Como se menciona anteriormente, el objetivo general de este estudio es el de diseñar y probar un sistema de detección de estrés en señales de voz.

La motivación que impulsa este objetivo se basa en la necesidad de diseñar un detector lo suficientemente preciso para situaciones reales de estrés, un aspecto clave como se introdujo antes. Dadas las aplicaciones que se han identificado para el uso del detector, la fiabilidad se convierte en una característica muy importante y, gracias a los avances tecnológicos que se han alcanzado en los últimos años, la idea de realizar este proyecto se hace posible.

Otras de las razones por las que se aspira a diseñar este sistema es porque su investigación podría enfocarse nuevos objetivos como: mejorar los sistemas de reconocimiento de voz actuales, o proporcionar sistemas que permitan ayudar al ser humano, mediante la detección de estrés, a tomar consciencia y corregir comportamientos en situaciones de estrés, o bien hacerlo por medio de un especialista.

Por otro lado, el proyecto se centrará en el cumplimiento de objetivos particulares que ayuden a alcanzar el propósito general.

En primer lugar, una de las intenciones de este estudio es, basándonos en la estructura de la base de datos que utilizaremos, encontrar el grado de relación que existe entre la voz, el estrés que sufre el locutor y su ritmo cardíaco. También resulta interesante la búsqueda de diferencias basadas en el género y en la edad de los participantes durante los periodos de estrés.

Otro objetivo sería responder a la pregunta de cómo afecta la calidad de las grabaciones de voz a la hora de detectar estrés a partir de sus características.

Así mismo, en el estudio se probarán diferentes máquinas para la detección con el fin de buscar qué sistemas proporcionan mejores resultados. Se experimentará también con varios conjuntos de datos tanto para el entrenamiento como para el test de la máquina.

Por último, el fin de este proyecto es contribuir a la literatura y a la tecnología actual haciendo públicos los desarrollos y resultados obtenidos[1].

1.3. Entorno socio-económico

El análisis de estrés en la voz, también conocido como *Stress Voice Analysis* (VSA), se hizo popular a mediados de los años 70, desarrollado para uso de Inteligencia Militar en busca de una alternativa al cable detector de mentiras.

A partir de entonces, los últimos avances en este campo han permitido ampliar el abanico de aplicaciones, sobre todo enfocadas al campo laboral y económico. Además de emplearse como detector de mentiras, la detección de estrés se utiliza hoy en día para realizar estudios de honestidad sobre las personas en situaciones como: pruebas de selección en empresas, eventos confidenciales, negociaciones de contratos y prevención de fraudes y de reclamo en aseguradoras.

Como se mencinó anteriormente, el VSA puede tener gran utilidad en aquellas profesiones en las que, bajo situaciones de estrés, se deben tomar decisiones rápidas. El desarrollo de sistemas de detección de estrés permitiría detectar cuándo el empleado ha dejado de ser eficiente y, por tanto, se evitaría correr riesgos causados por malas decisiones, lo que puede incluso salvar vidas.

En cuanto al campo de salud, el diseño de un detector de estrés supondría una ayuda a los especialistas para, por ejemplo, detectar características o indicadores de enfermedades, automatizar procesos clínicos o incluso evaluar si la situación de tensión del profesional es correcta para desempeñar su trabajo.

Por tanto, el entorno socio-económico actual dibuja un panorama en el que el desarrollo de estos sistemas de detección de estrés puede tener un gran impacto en una amplia variedad de profesiones, tanto como herramienta de ayuda para desempeñar el trabajo, como para evaluar si el trabajador se encuentra en las condiciones óptimas para ejercer.

[1] Más información en el Apéndice B

1.4. Marco regulador

Actualmente no se ha encontrado ninguna legislación aplicable exclusivamente a ninguna de las técnicas del habla, ni en concreto a la detección de estrés en señales de voz. En consecuencia, solo cabe destacar a nivel nacional la Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal (LOPD), en cuanto a los datos que se utilizan para estos estudios.

Por otro lado, debido a la fuerte relación que existe entre nuestro estudio y el campo de la ética, cabe mencionar que, a pesar de no haber una regulación específica en los aspectos que se refieren al proyecto, los organismos de financiación de la investigación sí establecen normas correspondientes a este ámbito [3].

Además, debido a que este campo constituye una rama de la Inteligencia Artificial, cabe mencionar la última publicación de octubre de 2016 de la Unión Europea disponible[4], enfocada en leyes sobre robótica.

Por último, en referencia a los datos utilizados para el estudio, la base de datos es pública y se encuentra disponible online[5]. Las muestras de voz se tomaron en eventos reales en los que se realizaban los discursos. La anotación se recogió mediante sensores fisiológicos no invasivos y fáciles de usar. Así mismo, los participantes del estudio fueron debidamente informados de las sesiones de grabación previas al evento y rellenaron los cuestionarios sobre salud voluntariamente, habiendo cumplimentado los formularios de consentimiento correspondientes como se indica en la literatura disponible [6].

1.5. Contenido de la memoria

A continuación se presentará de manera breve una descripción sobre el contenido de cada capítulo de la memoria con el fin de facilitar su comprensión.

El Capítulo 1 consiste en introducir al lector al contexto del proyecto, mostrar su entorno socio-económico, marco legal así como los principales objetivos que se van a cubrir.

El Capítulo 2 narra de manera detallada las tecnologías y estudios relacionados con nuestro tema del proyecto, dando al lector una visión global de los avances que se han alcanzado hasta la fecha, así como de los recursos de los que se disponen para su desarrollo.

En el Capítulo 3 se detalla de forma teórica la estructura de la solución del proyecto. También justificamos el diseño que vamos a seguir para este, es decir, elegimos algunas de las tecnologías y recursos explicados en el Capítulo 2 y justificamos por qué lo hemos elegido y el uso que le vamos a dar.

El objetivo del Capítulo 4 es detallar y explicar de manera minuciosa todos los procesos seguidos durante la implementación de proyecto, desde la obtención de la base de datos, hasta los resultados finales de clasificación.

En el Capítulo 5 se recogen una serie de pruebas, resultados y gráficas cuyo fin es evaluar los diferentes aspectos de nuestro sistema de detección con el fin de extraer conclusiones.

En el Capítulo 6 se realiza un análisis de la planificación del proyecto, desglosando las tareas y la ejecución temporal, y una estimación del presupuesto total necesario para el desarrollo del proyecto.

En el último Capítulo 7 se extraen las conclusiones del proyecto y se proponen líneas futuras para el mismo.

Los apéndices del final de la memoria tienen como objetivo aportar información adicional al proyecto. El Apéndice A recoge un desglose y detalle de los resultados del Capítulo 5. El Apéndice B contiene la información relacionada con el código desarrollado durante el proyecto. Por último, el Apéndice C incluye el correspondiente resumen en inglés del proyecto.

Capítulo 2

Estado del arte

2.1. Definición de estrés y tipos

Definir el estrés, más concretamente la voz bajo situaciones de estrés, resulta una tarea difícil de realizar de manera precisa. Según la definición propuesta por Murray et al. [7] , “el estrés es la variabilidad observable en ciertas características del habla debido a una respuesta a los factores de estrés". Por otra parte, Hansel [8] define en su libro el estrés reflejado en la voz como “un estado psicológico que es respuesta a una amenaza percibida o demanda de tareas y que, normalmente, se acompaña de emociones específicas”. De la misma manera, Hansel señala uno de los aspectos importantes sobre el estrés, que es la dificultad de separarlo del resto de las emociones, ya que normalmente surge como una combinación de ellas, como el enfado o la tristeza.

Estudios sobre psicología[9] establecen una clasificación del estrés en función de su duración:

- **Estrés agudo.** Es el estrés más común y ocurre a causa de las exigencias impuestas a uno mismo o sobre los demás. Este tipo de estrés es a corto plazo y puede presentarse en cualquier momento, como en una entrevista de trabajo. Este estrés suele ir acompañado de una sobreexcitación del sistema nervioso (sudoración, palpitaciones, aumento presión sanguínea...).
- **Estrés agudo episódico.** Aparece en las personas que sufren estrés agudo con mucha frecuencia debido a una gran cantidad de responsabilidades, cargas o situaciones de tensión.
- **Estrés crónico.** Es el estrés en su grado más alto. Suele tener graves consecuencias psicológicas e incluso físicas sobre el individuo.

En nuestro estudio, trataremos el estrés agudo así como su relación con el ritmo cardíaco y la forma en la que este se refleja en la voz del individuo.

2.2. Tecnologías del habla

Las tecnologías del habla son el conjunto de técnicas, métodos y algoritmos diseñados para modelar la comunicación hombre - máquina a partir del lenguaje oral (voz humana). Libros como *Speech and Audio Signal Processing*[10] y *Speech and language processing*[11] son algunos de los clásicos que se utilizan en el estudio de este campo. Las tecnologías del habla, como rama de la Inteligencia Artificial (AI), también poseen una fuerte relación con otras disciplinas como la lingüística, la fonética y la física. A continuación, se definen brevemente algunas de las tecnologías más utilizadas [12]:

- **Reconocimiento del habla (Speech Recognition, SR).** Consiste en la ciencia que se encarga de procesar la señal de voz del ser humano y de reconocer la información contenida en ella, esencialmente el texto. También se le conoce como Automatic Speech Recognition (ASR) o Speech-to-Text (STT). Los sistemas de reconocimiento de habla se pueden clasificar según su dependencia respecto al hablante. Esto significa que para diseñar y entrenar el modelo tendremos en cuenta o no las características de la voz de cada usuario de forma independiente.

- **Síntesis de habla (Speech Synthesis).** Se trata de la producción artificial de voz humana. Este sistema, también conocido como Text-To-Speech (TTS), convierte el texto normal en señal de voz, con el objetivo de hacer que se parezca lo máximo posible a la del ser humano y a nuestra forma de entenderlo. Por tanto, la síntesis de voz supone el proceso contrario al reconocimiento del habla.

- **Reconocimiento del hablante (Speaker Recognition).** Esta técnica consiste en la identificación de la persona que habla a partir de las características biométricas de su voz. En este punto, cabe destacar la diferencia que existe entre *Speaker Recognition*, es decir, el reconocimiento de *quién* habla, frente a *Speech Recognition*, que consiste en la identificación de lo que se está diciendo. Estos dos términos suelen ser confundidos con frecuencia, ya que ambos pueden ser denominados como *reconocimiento de voz*. Reconocer al hablante puede simplificar tareas como la de traducir discursos en sistemas que han sido entrenados con voces de personas específicas, o también para verificar la identidad de un hablante como parte de un proceso de seguridad.

- **Codificación de voz (Speech Coding).** La codificación de voz consiste en la compresión de datos de señales digitales de audio que contienen voz. Esta tecnología utiliza: la estimación de parámetros específicos del habla, mediante técnicas de procesamiento de señal de audio, para modelar la señal de voz y algoritmos genéricos de compresión de datos, para representar los parámetros modelados resultantes en un tren de bits compacto. Las dos aplicaciones más importantes en los que se utiliza la codificación de voz es la telefonía móvil y la voz sobre IP (VoIP).

2.3. Caracterización empírica del estrés

2.3.1. Entornos

Hasta la fecha, los estudios sobre reconocimiento de emociones o estrés han seguido varios caminos a la hora de obtener grabaciones de voz en diferentes situaciones de estrés. Algunas de los métodos más utilizados son:
Según el lugar donde se produzca la grabación:

- **Evento real.** Se toman grabaciones de los hablantes en situaciones reales de estrés, como por ejemplo, discursos en público.
- **Laboratorio.** El hablante es grabado en situaciones de estrés artificiales enfocadas al estudio de las mismas. Dentro de esta categoría, podemos distinguir según el tipo de discurso:[13]
  - **Interpretado.** Actores interpretan el estrés o las diferentes emociones que se estudien. Esto solía tener la ventaja de que resultaba fácil distinguir los diferentes estados de estrés o emociones. Sin embargo, no podían considerarse una aproximaición fiable [14].
  - **Inducido.** Los hablantes son sometidos a experiencias y factores de estrés por los investigadores.

Comúnmente se suele incluir en estas grabaciones la toma de medidas biométricas, detalladas más adelante en la Subsección 2.3.3, con el fin de ayudar a analizar las señales de voz, proporcionar más información para la detección de estrés e intentar buscar una relación entre los datos.

2.3.2. Bases de datos

Vervedis y Kotropoulos [13] recogen en su artículo una tabla con información exhaustiva de un total de 64 bases de datos existentes relacionadas con emociones, estrés y diversas medidas biométricas. Sin embargo, en esta sección se detallarán las bases de datos más relevantes para nuestro estudio sobre la detección de estrés:

- **SUSAS.** Esta base de datos[15] se recogió con el objetivo de formular y analizar algoritmos de habla bajo estrés y ruido. Una de las características importantes de las grabaciones es que se evalúa la voz bajo estrés tanto simulado como real. Por ello y por la gran variedad de datos que recoge es, generalmente, la base de datos más utilizada en el estudio de la variabilidad de la producción de voz en condiciones de estrés y tensión.

  La base de datos consta de un total de 44 hablantes (14 mujeres, 30 hombres), con edades comprendidas entre 22 y 76 años que generaron más de 16.000 expresiones. SUSAS divide sus datos en cinco dominios diferentes que incluyen (i) datos de análisis psiquiátrico (habla bajo depresión, temor, ansiedad), (ii) diferentes estilos de conversación (enfado, calma, rápido, alto, lento, bajo), (iii) grabaciones durante tareas simples, altas

cargas de trabajo y situaciones de ruido (efecto Lombard), (iv) archivos con doble respuesta entre el ordenador y el usuario, y (v) audios durante la realización de tareas para combatir miedos (Gforce, efecto Lombard, ruido). La base de datos tiene en total un vocabulario 35 palabras comunes en el ámbito de la comunicación de aeronaves.

- **VOCE Corpus.** Se trata de la base de datos que utilizaremos en nuestro estudio. A diferencia de la literatura que se encuentra disponible [16], la versión actualizada [5] de VOCE Corpus recoge un total de 135 grabaciones de voz que resultan de un conjunto de 45 estudiantes (21 hombres, 17 mujeres y 7 no identificados) de la Universidad de Porto, con edades de entre 19 y 49 años.

  Estos archivos de voz (formato .wav) corresponden a tres momentos diferentes de grabación: *Prebaseline*, lectura de un texto estándar un mínimo 24 horas antes del evento en público (Public Speaking, PS); *Baseline*, lectura del mismo texto que en *Prebaseline* aproximadamente 30 minutos antes del PS; y *Recording*, grabación del evento en público de texto y duración libre.

  Junto a estos archivos de voz, se recogen 117 archivos (formato .xml), cada uno de ellos correspondiente a un archivo de voz, que contienen 2 medidas para evaluar el ritmo cardíaco (Heart Rate, HR). Estas medidas, tomadas con el aparato Zephyr HxM BT2 [1], son (i) valores denominados *Zecg* que representan un valor de HR promediado y filtrado y (ii) valores *Zts* que hacen referencia a los instantes de tiempo en los que se producen picos R[2] en el electrocardiograma obtenido con el aparato, medidos con un reloj interno de 16 bits. Cada uno de estos valores va acompañado del instante de Tiempo Universal Coordinado (Universal Time Coordinated, UTC) correspondiente.

  Adicionalmente, la base de datos cuenta con un archivo *metadata* (formato xml) que recoge, para 38 de los 45 individuos, datos sobre género, edad, información sobre salud, sobre experiencia al hablar en público, puntuaciones de test STAI[3] e información sobre la calidad de las grabaciones (nivel energía, saturación...).

  En total, la base de datos reúne información completa[4] de 33 individuos y 12 incompletos.

[1] Información disponible en https://www.zephyranywhere.com/media/download/hxm1-api-p-bluetooth-hxm-api-guide-20100722-v01.pdf

[2] Más información en la Subsección 2.3.3

[3] Más información en la Subsección 2.3.3

[4] En este estudio consideraremos que la información es completa cuando un usuario tenga los 3 archivos de audio y sus correspondientes archivos de HR

- **Otras bases de datos utilizadas en algunas de las referencias mencionadas en este estudio son:**

  La base de datos de Fernández y Picard [17], nombrada por Ververidis y Kotropoulos [18] como "Base de Datos 11", como se describe en el artículo, está compuesta por 598 expresiones, correspondientes a cuatro hablantes diferentes. Se les pidió a los sujetos que respondieran preguntas matemáticas mientras conducían camiones a dos velocidades diferentes: 60 mph y 120 mph. Mientras que a la primera velocidad se pedía al sujeto una respuesta cada 9 segundos, a la velocidad de 120 mph se solicitaba respuesta cada 4 segundos.

  Por otro lado, en el artículo de Demenko [19] [20] se utiliza una base de datos que recoge llamadas de emergencia de la Policía de Poznan, formada por grabaciones de voz espontánea que consta de notificaciones de delito y solicitudes de intervención policial. Cuarenta y cinco usuarios fueron elegidos después de una previa evaluación acústica. La anotación preliminar fue realizada por estudiantes formados en fonética, e incluyeron descripción del tipo de diálogo, características del discurso (características suprasegmentales), contexto y acústica de fondo.

2.3.3. Anotación del estrés y biometría

Para realizar la anotación del estrés en la voz, la literatura presenta diversas medidas tanto psicológicas como biomédicas, que, combinadas con las características extraídas de la voz como explicaremos en la siguiente sección, constituyen los datos necesarios para detectar si en una señal de voz el hablante presenta estrés o no.

El ritmo cardíaco y los parámetros que se extraen de él son algunos de los factores más considerados a la hora de evaluar el estrés mediante biometría. Estudios como el de Mariana Dimas [21] utilizan los valores de los instantes en los que se producen los picos de HR (ondas R) y los intervalos entre ellos (intervalos RR) (Figura 2.1) para determinar la probabilidad de que exista estrés o no en los segmentos de la señal de voz. Del mismo modo, los valores de HR pueden variar según las edades y géneros, por lo tanto, esto supone un indicador más a tener en cuenta a la hora de evaluar el estrés.

Otros estudios [23][24] analizan la respuesta galvánica de la piel (Galvanic Skin Response, GSR) junto con las características de la voz para determinar esta posibilidad de estrés. La GSR, también conocida como respuesta electrodérmica (Electrodermal Response, EDR) o conductividad de la piel (Skin Conductance Response, SCR), es uno de los marcadores para detectar la activación emocional. La GSR se basa en el cambio de calor, en la electricidad que transmiten los nervios y en el sudor a través de la piel, todo ello debido al carácter eléctrico de la misma.

Ki-Seung Lee [25], por otra parte, realiza una publicación sobre el reconocimiento de habla basado en la electromiografía (EMG), es decir, el estudio de la actividad eléctrica de los músculos del esqueleto.

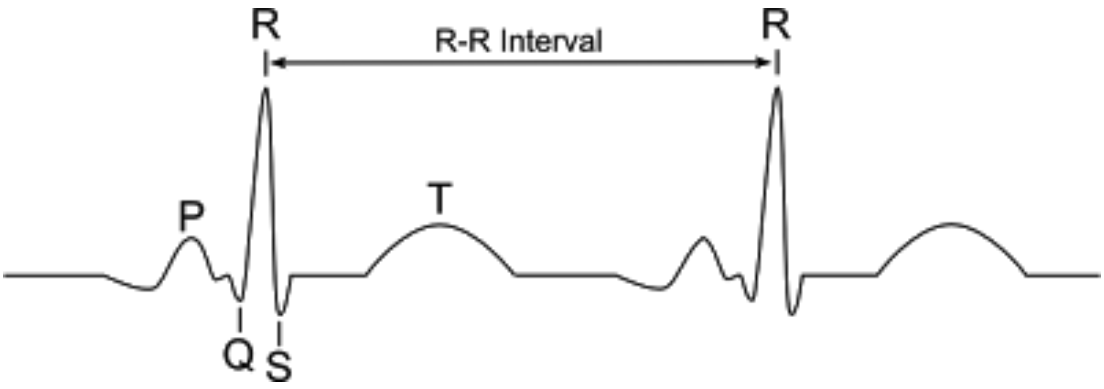

FIGURA 2.1: Señal de electrocardiograma (ECG) con ondas R e intervalos RR [22]

Podemos encontrar numerosas bases de datos en la literatura en las que se recogen otras medidas como la presión sanguínea, análisis de la sangre o datos sobre respiración como en las bases de Rahurkar and Hansen y SUSC-0, de Hansen[13].

No solo existen medidas biométricas para evaluar el estrés, sino también psicológicas. Para ello, se utilizan formularios de autoevaluación, como el Cuestionario de Ansiedad Estado-Rasgo (State-Trait Anxiety Inventory, STAI)[5] empleado para autoevaluar dos conceptos independientes de la ansiedad: estado (E), que evalúa la ansiedad en el instante en el que se realiza el test; y rasgo(R), que sirve para evaluar la ansiedad en general en una persona [26].

2.4. Métodos de extracción de características

Los métodos de extracción de características son aquellos algoritmos y técnicas cuyo objetivo es calcular un conjunto de vectores de características. Estos vectores proporcionan una representación compacta de aquellos aspectos más importantes de los datos de entrada al sistema, en nuestro caso, de las señales de voz.

La extracción de características es la fase principal de cualquier sistema de habla, sobre todo en los sistemas de reconocimiento. En esta fase se utilizan distintos métodos de extracción que juegan un papel importante a la hora de separar un discurso de voz de otro. Esto se debe a que cada voz tiene diferentes propiedades individuales contenidas en las expresiones o segmentos de la señal .

Características como las frecuencias de los formantes, el pitch (frecuencia fundamental), la duración, los Coeficientes Cepstrales en las Frecuencias de Mel (Mel Frequency Cepstral Coeffcients, MFCC) o el espectro del tracto vocal, son algunas de las propiedades que podemos obtener. A continuación se detallarán algunas de las técnicas más utilizadas actualmente así como una breve explicación de las características que se pueden obtener de ellas.

[5]Disponible en https://goo.gl/H4ZUWr

- **Codificación Predictiva Lineal (Linear Predictive Coding, LPC)**

  LPC es una las técnicas de análisis de señal más potentes entre los algoritmos lineales. Esta ténica se ha convertido en la predominante debido a que proporciona una estimación exacta de los parámetros del habla además de ser un modelo computacionalmente eficiente.

  La idea principal en la que se basa LPC es que una muestra de voz puede ser aproximada como una combinación lineal de muestras de voz anteriores. Mediante la minimización de la suma de cuadrados (sobre un intervalo finito) entre las muestras de voz reales y los valores predichos, se puede determinar un conjunto único de parámetros o coeficientes de predicción. Los formantes son una de las características que se pueden extraer mediante LPC.

  En acústica, los formantes se refieren a los picos en la envolvente del espectro del sonido y corresponden con las frecuencias de resonancia del tracto vocal. Así, el tracto vocal se modela como un filtro lineal con resonancias o frecuencias formantes.

  En Figura 2.2 se muestran gráficamente una trama de voz en el dominio del tiempo (arriba) y la envolvente espectral estimada mediante predicción lineal (abajo) cuyos picos se sitúan en las frecuencias de los formantes.

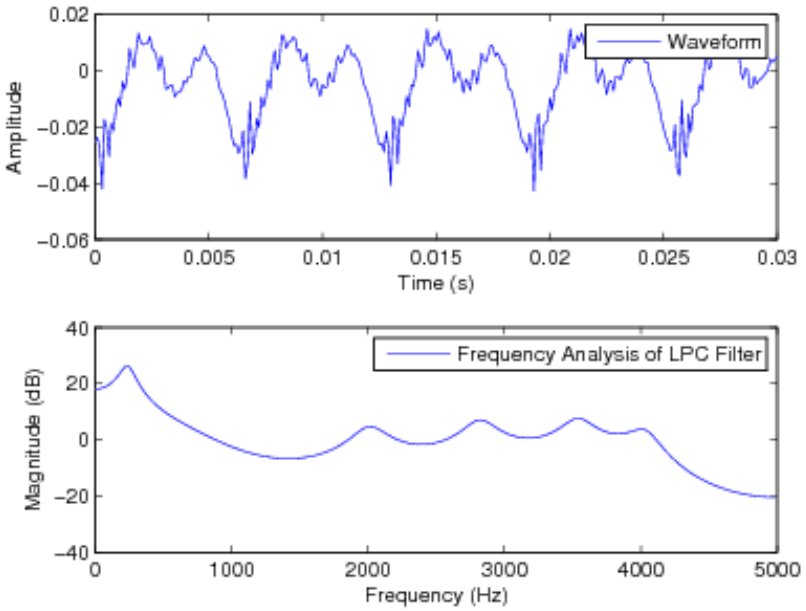

FIGURA 2.2: Representación gráfica de una trama de voz (arriba) y de su envolvente espectral (abajo) con las frecuencias de los formantes [27]

- **Mel Frequency Cepstral Coefficient (MFCC)**

  Los MFCC son coeficientes que se extraen de la voz basándose en algunos aspectos de la percepción auditiva humana. Estos se calculan a partir de la Transformada Discreta de Coseno (DCT) del logaritmo del espectro de la trama de audio al que se le ha aplicado una transformación no lineal del eje de la frecuencia. El proceso MFCC se lleva a cabo mediante cinco fases, como se muestra en la Figura 2.3.

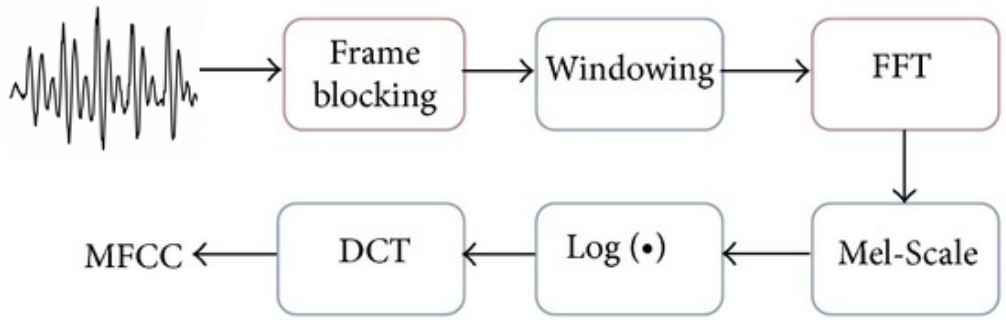

FIGURA 2.3: Diagrama de bloques del proceso de extracción de los MFCC [28]

  En el bloque de segmentación en tramas (Frame Blocking), la señal de voz se divide en tramas de aproximadamente 20-30 milisegundos. En la fase de enventanado (Windowing) se minimizan las discontinuidades presentes en la señal suavizando el comienzo y el final de cada trama. En el siguiente paso, el cálculo de la Transformada de Fourier (FFT) se utiliza para la conversión de cada trama del dominio de tiempo al dominio de frecuencia. Posteriormente, la señal se representa en el la escala de Mel (Melspectrum) para imitar el modelo del oído humano, cuya escala se parece más a la conocida como Mel (espaciamiento lineal por debajo de 1000 Hz y un escalamiento logarítmico por encima de 1000 Hz) en lugar de a la lineal. Seguidamente, se calcula el logaritmo y se realiza la Transformada Discreta del Coseno (Discrete Cosine Transform, DCT), de la que habitualmente sólo se toman los 12-13 primeros coeficientes obteniendo los MFCC.

- **Teager Energy Operator, TEO.** Las técnicas de extracción lineales se derivan de modelos de producción de voz que asumen que el flujo de aire que se propaga en el tracto vocal se modela como una onda plana [29]. Sin embargo, el sistema de voz tiene características no lineales que no son capturadas por estos sistemas lineales. Es por esto por lo que aparecen los TEO, con el objetivo de cruzar esta limitación. Estos operadores, en esencia, a diferencia de los lineales, no miden la energía de la propia señal, sino la energía que genera la señal en el sistema de habla.

- Otras de las características comunes que se suelen extraer en los sistemas de voz son **el jitter y el shimmer**. El jitter se define como el parámetro de variación de frecuencia de ciclo a ciclo, mientras que shimmer se refiere a la variación de amplitud de la onda sonora [30]. En la Figura 2.4 se representan gráficamente estas medidas.

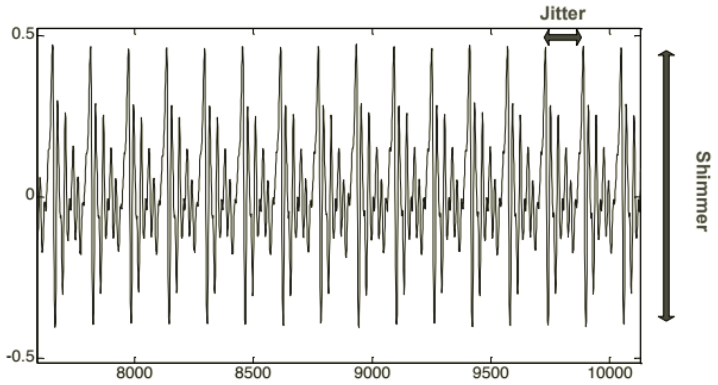

FIGURA 2.4: Representación gráfica del jitter y shimmer para una señal de voz [30]

- Por último, la **velocidad de habla o *Speech Rate*** a la que un locutor pronuncia una palabra u oración se utiliza como una caracterísca muy efectiva en el reconocimiento de emociones [31].

2.5. Clasificadores

El primer aspecto a tener en cuenta cuando hablamos de clasificadores son los dos grandes grupos que existen: los sistemas de clasificación supervisados y no supervisados. Los sistemas de clasificación supervisados son aquellos en los que, a partir de un conjunto de muestras ya clasificadas (conjunto de entrenamiento), intentamos asignar una clasificación a un segundo conjunto de muestras desconocidas. Por otra parte, los sistemas de clasificación no supervisados son aquellos en los que no disponemos de una batería de ejemplos previamente clasificados, sino que intentamos buscar una estructura o modelar los datos de los que disponemos a partir de las propiedades de los mismos.

En nuestro proyecto nos centraremos en los algoritmos de aprendizaje supervisado puesto que en general obtienen mejores resultados. A continuación se detallarán algunos de los algoritmos de clasificación más utilizados en las tecnologías del habla:

- **Modelos ocultos de Markov (Hidden Markov Models, HMM) [32].** En este clasificador la señal de voz es modelada como un conjunto de unidades acústicas que pueden ser consideradas como sonidos elementales del lenguaje. Tradicionalmente, la unidad elegida es el fonema, por tanto, una palabra estaría formada por fonemas concatenados. También se

pueden considerar otro tipo de unidades como sílabas, disílabos o fonemas en su contexto, lo que hace que el modelo sea más discriminativo. Sin embargo, según Schwartz[33] esta mejora teórica está limitada en la práctica por la complejidad involucrada y los problemas de estimación.

En los modelos HMM, la señal de voz trata como una serie de unidades acústicas que, típicamente, son modeladas por un autómata de estados finito con una topología de izquierda-a-derecha como se refleja en la Figura 2.5.

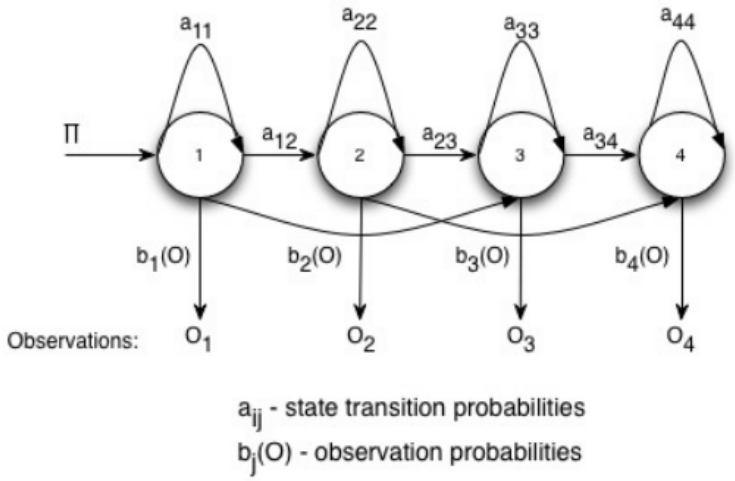

FIGURA 2.5: Diagrama de un HMM de izquierda a derecha[34]

- **Perceptrón multicapa (Multi-layer Perceptron, MLP.)**

  El Percepción Multicapa (MLP) es quizás la arquitectura de red más popular tanto para la clasificación como para la regresión. MLP es una red neuronal artificial (Artificial Neural Network, ANN) que suelen estar compuesta de varias capas de nodos con conexiones unidireccionales, a menudo entrenados por *backpropagation* (retropropagación). El proceso de aprendizaje de la red MLP[35] consta de las muestras de datos que forman el vector de entrada N-dimensional x y el vector de salida deseado M-dimensional d, llamado destino u objetivo (*target*). Procesando el vector de entrada x, el MLP produce el vector de señal de salida $y(x, w)$ más cercano al deseado mediante el ajuste del vector de pesos adaptados w, como se refleja en la Figura 2.6.

  El algoritmo de aprendizaje de MLP se basa en la minimización de la función de error definida en el conjunto de aprendizaje (x_i, d_i) para $i = 1, 2, 3, ...N$ utilizando, por ejemplo, la norma Euclídea:

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} \|y(x_i, w) - d_i\|^2$$

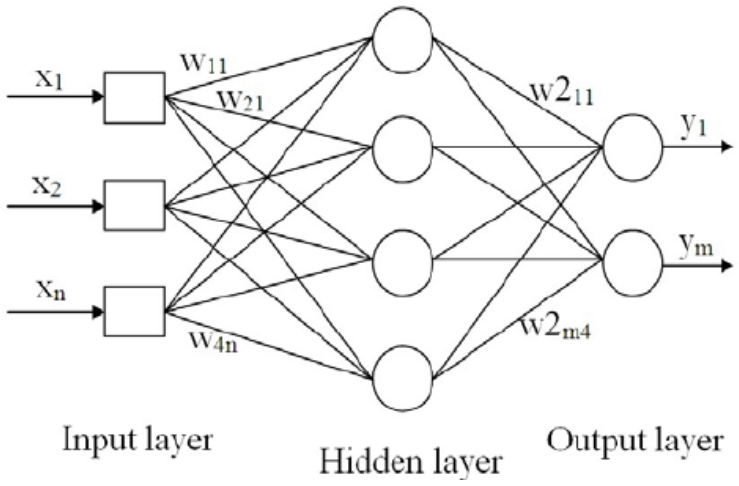

FIGURA 2.6: Diagrama de un modelo de MLP de una capa oculta[36]

La minimización de este error conduce a los valores óptimos de los pesos. Los métodos más eficaces de minimización suelen ser los algoritmos de descenso según el gradiente.

- **Máquina de Vector Soporte (Support Vector Machine, SVM)**

  Los clasificadores SVM están basados principalmente en el uso de funciones no lineales que mapean las características originales en una dimensión del espacio mayor, permitiendo así la clasificación mediante un clasificador lineal.

  Sin embargo, no siempre se puede hacer la clasificación mediante una linea recta. Los SVM utilizan distintas funciones Kernel para tratar casos más complejos como en los que aparecen más de dos variables predictoras, curvas no lineales de separación, conjuntos de datos que no pueden ser completamente separados o clasificaciones multiclase. Además de la función lineal, existen otras funciones Kernel para realizar estos casos como la polinomial-homogénea, la función de base radial (Radial Basis Function, RBF) o la sigmoide.

2.6. Estudios relacionados

Tras analizar las diferentes fases de los sistemas de habla, a continuación, con el fin de entender el concepto del sistema completo, se detallarán algunos de los estudios que han desarrollado sistemas de detección de estrés. En la figura Figura 2.7 se representa el esquema general que suelen seguir estos sistemas.

El estudio de O. Simantiraki [38] propone un conjunto poco habitual de características basadas en la inclinación espectral de la fuente glotal, además de

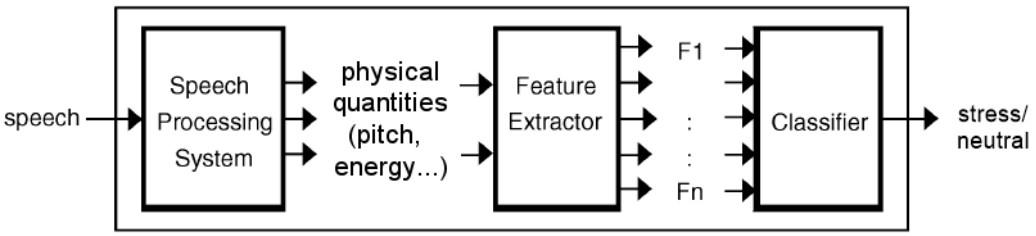

FIGURA 2.7: Diagrama de bloques que representa la estructura de un sistema de detección en voz [37]

las de la propia señal de voz. Las características se calculan a partir de la función de densidad de probabilidad de las pendientes espectrales estimadas, y consisten en las tres pendientes más probables de la fuente glotal, así como las correspondientes tres pendientes de la señal de voz, obtenidas a nivel de palabra. El rendimiento del método propuesto se evalúa en el conjunto de datos simulado de la base de datos SUSAS, logrando una precisión de reconocimiento del 92,06 % mediante un clasificador Random Forest basado en árboles de decisión.

Otro ejemplo interesante es el proyecto que realizaron en la Universidad de Alcalá de Henares sobre la detección de estrés a través del análisis de las emociones en el habla[39]. En el estudio utilizaron la base de datos EMO-DB (Berlin Database of Emotional Speech) para extraer un conjunto de características estándar: MFCC, energía, pitch, jitter, shimmer y la relación armónico-ruido, para medir la pureza de la voz. La clasificación del estrés la realizan a través de la agrupación de 7 emociones en grupos de bajo, medio y alto contenido en estrés. Los resultados obtenidos para la clasificación de 3 clases resultó entorno al 80-85 % de acierto, mientras que para la de 7 clases, obtuvieron un rendimiento de aproximadamente 63-68 %.

Por último, cabe mencionar el estudio de Mariana Dimas de la Universidad de Porto[21], cuyo objetivo fue encontrar el set de características más reveladoras para la detección de estrés en voz. Para ello, propusieron extraer un set inicial de 6365 características, entre las que incluyen tanto características funcionales como TEO. Posteriormente, se realizó una fase de selección de características, obteniendo un set reducido como entrada para la clasificación binaria entre estrés y neutral. Para este último paso, se utilizan clasificadores SVM cuyo rendimiento resultó en torno al 65-75 %.

Capítulo 3

Diseño

3.1. Descripción del problema

Como se ha planteado en los capítulos anteriores, y como dice el propio título del proyecto, el problema que trataremos de resolver es el de la detección de estrés en señales de voz, es decir, conseguir determinar si, tanto de forma global o por segmentos de la señal de voz, el individuo presenta estrés o no.

Para realizar esto, utilizaremos la base de datos VOCE detallada anteriormente en la Subsección 2.3.2. Contaremos, por tanto, con un conjunto de grabaciones de voz así como de datos fisiológicos, en nuestro caso, valores relacionados con el ritmo cardíaco tomados durante los tiempos de grabación.

Antes de pasar a la fase de extracción de características, será necesario analizar la base de datos, es decir, examinar la calidad de sus audios y detectar los archivos de los que dispone cada uno de los usuarios que participan en las grabaciones. Este análisis servirá para formar diferentes conjuntos con los que se realizarán las correspondientes pruebas.

Una vez realizado esto, se extraerán las características de los audios y se generarán las etiquetas o *target* a partir de sus datos fisiológicos. Para esto último, será necesario determinar los umbrales de decisión con los que se clasificará cada segmento o trama de la señal de voz con un 1 (estrés) o un 0 (no estrés).

Por último, una vez tengamos toda la información anterior, el objetivo será diseñar varios sistemas o máquinas de detección con los algoritmos detallados en la Sección 2.5. La finalidad de esto será, tras probar diferentes parámetros en las máquinas, evaluar los resultados y extraer las conclusiones correspondientes.

3.2. Estructura y diseño de la solución

En esta sección se detallarán mediante diagramas de bloques los pasos y la estructura de la solución propuesta para resolver el problema que se plantea.

En primer lugar, como se muestra en la Figura 3.1, se realizará la extracción de características a partir de las grabaciones de audio, así como la generación de las etiquetas a partir de los datos de HR. Ambos procesos se realizarán a dos niveles o tamaños de segmento, señalado en el paso (1), obteniendo así las *features1* y *labels1*, con un tamaño de segmento menor, junto a lo que denominaremos *features2* y *labels2*, con un tamaño de ventana mayor. Además de esto, para el caso de las etiquetas se diseñarán dos umbrales de decisión, como refleja el paso (2), con los que se decidirá si hay estrés o no. El primer umbral de decisión se basará en estadísticos de los valores de HR, obteniendo así *labels11* y *labels12*, mientras que el segundo umbral se tomará a partir de un percentil, dando lugar a *labels21* y *labels22*. Este proceso se detallará de forma más precisa en el capítulo Capítulo 4.

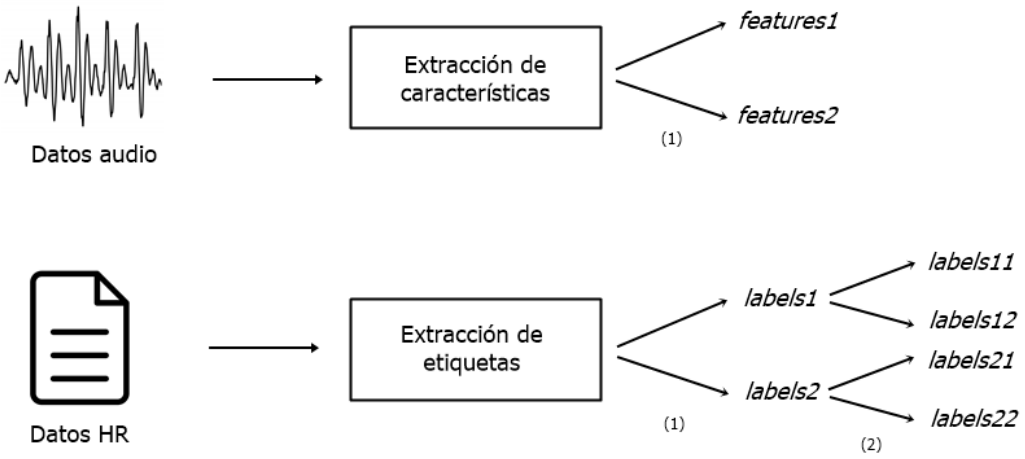

FIGURA 3.1: Diagrama de bloques del proceso de extracción de características y la generación de etiquetas

Tras obtener estos datos, pasaremos a la fase del diseño de la máquina y clasificación. Para ello, como es habitual en estos problemas, dividiremos nuestros datos, es decir, el set de características y etiquetas por segmento, y los dividiremos en conjuntos de entrenamiento (80 %) y de test (20 %). En la Figura 3.2 se explica el proceso completo con detalle.

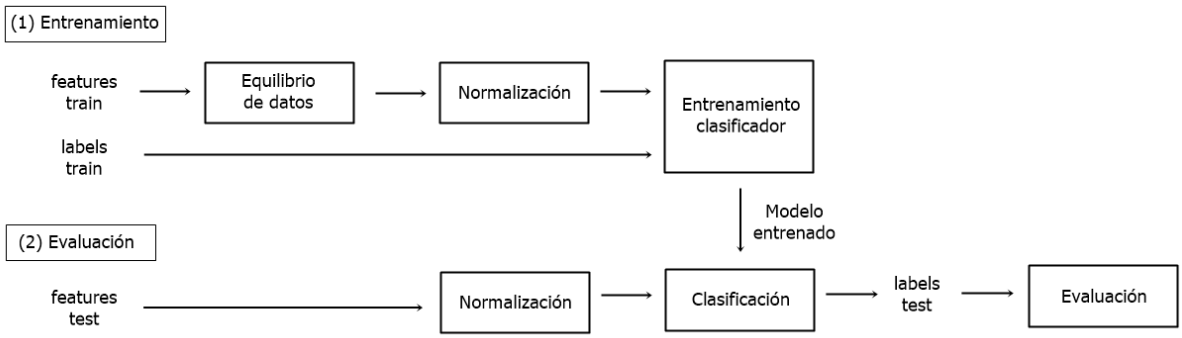

FIGURA 3.2: Diagrama de bloques del proceso de entrenamiento, clasificación y evaluación del sistema

3.3. Justificación elección base de datos

Tras analizar las bases de datos disponibles en este campo en la Subsección 2.3.2, se ha decidido utilizar la base de datos VOCE por el motivo principal de que, tanto las grabaciones de voz como los datos fisiológicos, se tomaron en eventos de estrés reales. Este motivo es de suma importancia, como se explicó en en Capítulo 1, ya que nos permite llegar a resultados aplicables a la realidad, a diferencia de los resultados obtenidos a partir de datos de laboratorio.

Además de esto, elegimos este conjunto de datos debido a los diferentes periodos de grabación que presenta: *prebaseline, baseline y recording*. Esto será la base para determinar los umbrales de decisión estrés/no estrés, ya que, tras analizar los diferentes archivos, se observó que, en la mayoría de casos, los momentos de *prebaseline y baseline* servirían como punto de partida o estado neutro de los participantes, y, el momento de *recording*, aquél en el que el usuario se encontraba a un nivel de estrés superior.

Por último, otro punto interesante por el que se ha decidido utilizar esta base de datos ha sido por los datos personales de los pacientes, en concreto, los de género y edad, que permitirán realizar análisis concretos según estos parámetros.

3.4. Diseño de extracción de características

Como se ha detallado en la Sección 3.1, la extracción de características se realizará a dos niveles de segmento, es decir, dos tamaños de ventana con las que se analizará la señal, de los cuales el mayor se podrá calcular a partir del menor, como se detalla en la Figura 3.3.

Para este proceso se utilizarán tanto técnicas LPC como de extracción de los MFCC, como se han detallado en el Capítulo 2, basadas en algunos estudios sobre reconocimiento de emociones [40][41][42]. De esta forma, obtendremos características básicas como el pitch, las frecuencias de los formantes, los coeficientes MFCC y la energía de la señal.

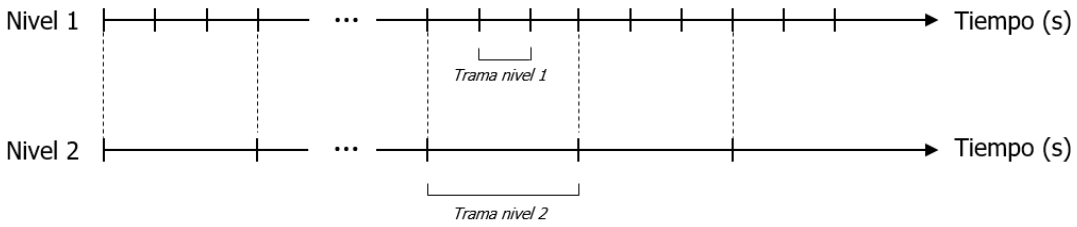

FIGURA 3.3: Niveles de extracción de características para la señal de voz

3.5. Elección de clasificadores

Para la fase de clasificación, se diseñarán máquinas basadas en los algoritmos de aprendizaje supervisado MLP y SVM, explicados en la Sección 2.5. Por tanto, generaremos un modelo para nuestras muestras a partir de un conjunto de entrenamiento y lo evaluaremos con los datos de test.

Los motivos por los que se han seleccionado estos algoritmos son, además de los numerosos estudios que encontramos en la literatura, las ventajas que nos pueden ofrecer en nuestro proyecto.

Las SVM, por ejemplo, son clasificadores que normalmente se utilizan en los casos binarios, como es el nuestro. Otras ventajas que presenta para nuestra clasificación es su versatilidad debido a las diferentes funciones Kernel que puede utilizar para la toma de decisiones. Además, las SVM resultan más eficientes cuando el conjunto de datos es relativamente pequeño, ya que es extremadamente robusto para la generalización.

Por otro lado, aunque el entrenamiento de un MLP suele ser mucho más costoso, los resultados de clasificación son mucho más eficientes, lo cual es un punto clave. Este clasificador es uno de los más usados de forma clásica debido también a su eficiencia sobre modelos no lineales.

Capítulo 4

Implementación

4.1. Procesado de la base de datos

Esta primera fase del proyecto es de gran importancia debido a que, al tratarse de una base de datos desconocida, resulta esencial la preparación y correcta interpretación de los datos para usarlos posteriormente al diseñar la solución de nuestro problema. A continuación se detalla de forma ordenada los pasos para realizar el procesado de la base de datos.

En primer lugar, la base de datos la forman grabaciones y archivos con valores de HR correspondientes a diferentes participantes. Cada uno de ellos estará identificado por un número único denominado *ID*. En la base de datos original algunos de estos ID eran números negativos y se tomó la decisión de hacer que todos fueran positivos para evitar futuras confusiones.

Otra de las grandes diferencias que encontramos en la base de datos respecto al artículo que la define[16], es que los nombres de los diferentes instantes de grabación no coinciden con los descritos. En la base de datos encontramos los instantes *prebaseline, baseline y recording*, mientras que en la literatura aparecen denominados como *baseline, experiment y event*. Tras analizar los UTC de los diferentes archivos y basándonos en los detalles de cuándo se tomó cada uno, asumimos que se cumple la siguiente correspondencia *base de datos* $\leftrightarrow$ *paper*:

- *prebaseline*$\leftrightarrow$ *baseline*
- *baseline* $\leftrightarrow$ *experiment*
- *recording*$\leftrightarrow$ *event*

A continuación, se realizó un listado de todos los archivos disponibles en la base de datos. De esta forma, se observó que para el total de 45 IDs que la forman, 33 de ellos cuentan con los tres archivos de audio y sus correspondientes datos de HR, mientras que 12 de ellos estaban incompletos.

Una vez disponíamos de esta información, comenzó la extracción de datos de los archivos xml que contienen los valores relacionados con el HR. Esta extracción se realizó con la herramienta MATLAB y sus funciones para leer archivos xml. Como se ha descrito en Subsección 2.3.2, estos archivos estan formados por dos valores distintos: *Zecg* y *Zts*. Recordamos que *Zecg* son valores promedios de HR, mientras que *Zts* se denominaba a los instantes de tiempo en los que se producen picos en el electrocardiograma.

Cada uno de estos valores se ha analizado y transformando por separado en valores que posteriormente se pudieran utilizar en nuestro diseño:

- En cuantos a los valores de Zecg que se obtienen directamente de los archivos, y asumiendo que la unidad en la que estan es en pulsos por minuto *(beats per minute, bpm)* se observaron dos características: (i) las muestras estan recogidas con una resolución de, aproximadamente, 1 muestra por segundo, y (ii) algunos de los valores Zecg eran negativos, lo cual no tenía sentido para valores de ritmo cardíaco.

  Tras analizar los manuales del aparato Zephyr HxM BT2 [1] con el que se tomaron los valores, se observó que estos debían ser de tipo *unsigned*, es decir, valores de 0 a 255, mientras que los valores obtenidos estaban en un rango de -128 a 127, es decir, como tipo *signed*. Tras esta apreciación, se transformaron todos los valores al rango de 0 a 255.

- Por otro lado, los datos de Zts tuvieron que ser procesados de manera minuciosa debido a la dificultad de su interpretación. El primer inconveniente que encontramos fue que no todos los archivos contenían los valores Zts, por lo que limitó aún más la información de nuestra base de datos. Para los que sí los tenían, observamos dos cuestiones que resolver: en primer lugar, los valores estaban tomados con un reloj de 16 bits que se reiniciaba, por tanto, no teníamos los valores de tiempo reales; en segundo lugar, en algunos instantes UTC aparecían varios valores de Zts de forma desordenada, es decir, aparecían en primer lugar el que había ocurrido más tarde y viceversa. Ambos casos podían aparecer tanto individualmente como juntos.

  Tras solucionar estos problemas de formato, el último paso fue pasar los valores de Zts a unidades *bpm* para poder compararlos con Zecg y ver la relación entre ellos. En la Figura 4.1 podemos observar dicha comparativa.

Como comprobamos en la Figura 4.1, los valores de Zecg son, efectivamente, un valor promedio y filtrado del HR original, representado con los valores de Zts procesados. Esta representación nos sirvió para tomar la decisión de utilizar, a partir de este punto del proyecto, únicamente los valores Zecg, asumiendo la pérdida de precisión que lograríamos con los valores Zts, pero solucionando el problema para los archivos que no contenían dichos valores.

Estos valores de HR denominados Zecg serán de gran utilidad en el proyecto para dos cuestiones clave: en primer lugar, se utilizarán como criterio para la creación de sets de datos y, en segundo lugar, serán la base para generar las etiquetas de las muestras de audio.

[1] Información disponible en https://www.zephyranywhere.com/media/download/hxm1-api-p-bluetooth-hxm-api-guide-20100722-v01.pdf

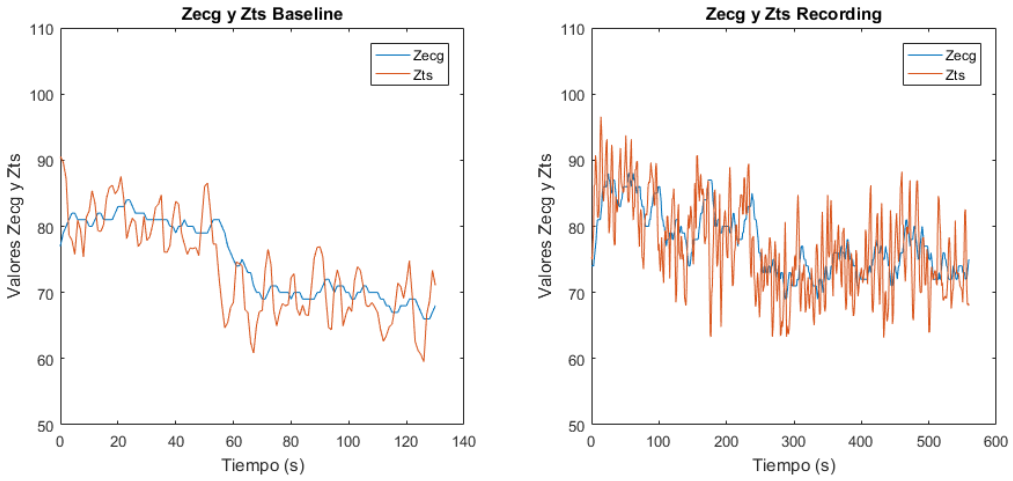

FIGURA 4.1: Comparativa de los valores Zecg y Zts para los archivos *baseline* y *recording* del ID 12782919

4.2. Selección de datos para el proyecto

La creación de sets de datos, es decir, la selección de audios y archivos de los diferentes IDs que se utilizaran en el proyecto supone una fase clave en el proceso de implementación ya que de ello dependen los resultados finales.

De forma inicial, y tras escuchar de forma manual todos los audios de la base de datos, se hizo un primer barrido del conjunto de IDs, descartando aquellos que cumpliesen alguno de los siguientes criterios:

- **Criterio 1:** Falta de, al menos, el archivo de HR del instante *recording*.
- **Criterio 2:** Audios, al menos del instante *recording*, vacíos o que contengan únicamente ruido de fondo.
- **Criterio 3:** Aquellos IDs cuyas grabaciones de los 3 instantes fueran repetidas.
- **Criterio 4:** IDs cuyos audios tuvieran calidad realmente mala como para no poder diferenciar el discurso del ruido.

En la Tabla 4.1 se recoge el total de los 21 IDs descartados según el criterio que le corresponde.

Una vez obtenido el conjunto completo de IDs útiles para nuestro proyecto, se procedió a la creación de distintos sets para utilizar en el diseño de la máquina y en la evaluación de los resultados.

En este proyecto se propone la creación de 2 sets de datos: el *Set 1*, formando por los IDs cuyos audios son de mayor calidad, longitud relativamente corta, y valores medios de Zecg más lógicos, es decir, aquellos con menor desviación típica y cuya media *recording* es superior a la media *prebaseline o baseline*. Los IDs que no cumplen estos criterios forman el denominado *Set 2*.

Criterio 1		Criterio 2	Criterio 3	Criterio 4	
286484722	1499703648	105804962	1507583907	192369217	548414142
550155379	1884865801	348334269		334342001	724250529
638882617	2071492831	1610132012		354408438	928432217
1062237033	2084620463	1650434878			
1113564542	1411484167				

CUADRO 4.1: Conjuntos de IDs agrupados según el criterio que descartó su uso en el proyecto

En este punto del desarrollo se plantea un nuevo problema ya que, de forma general y tras el análisis de archivos, los instantes *prebaseline* y *baseline* resultan similares comparándolos con el instante *recording*, representando ambos la situación de reposo del hablante. Por tanto, decidimos elegir, para cada ID, sólo uno de estos dos archivos. El criterio para tomar esta decisión fue elegir aquel archivo cuya desviación típica de los valores Zecg, calculada durante todo el periodo de grabación del audio correspondiente, fuera menor. Este criterio se tomó basándonos en la idea de que, si estos instantes significaban reposo, queríamos que la variabilidad de HR sea la menor posible.

Por tanto, en resumen, en este punto del proyecto tendremos, para cada ID del set correspondiente: dos grabaciones, una del instante *recording* y otra del instante que a partir de ahora denominaremos *base*, que corresponde al que se ha elegido entre *prebaseline* o *baseline*; y tendremos además los valores Zecg, es decir, de HR, correspondientes a cada archivo de audio.

Set 1		Set 2	
62963719	1397020749	12782919	902398068
652033332	1420900415	49425811	1143102813
935941053	1739028311	92305089	1458206716
1015666824	1777769661	304102792	1626125349
1395228143	2054751935	334844205	1686645257
		513604950	1756953694
		852630991	1777108864

CUADRO 4.2: Conjuntos de datos set 1 y set 2 agrupados según el ID

En la Tabla 4.2 se presenta tanto el set 1 (compuesto de 10 IDs) como el set 2 (14 IDs) junto con sus correspondientes identificadores que los forman.

Por último, denominaremos como *Set 3* el conjunto formado por la unión de Set 1 y Set 2 para las futuras pruebas del Capítulo 5.

4.3. Preprocesado de señales de voz

La fase de preprocesado de las grabaciones de voz tiene dos objetivos principales: normalizar todas las señales para poder compararlas entre sí y prepararlas para realizar posteriormente la extracción de características.

A continuación, se representa en la figura 4.2 el proceso que seguirán los archivos de audio durante este proceso:

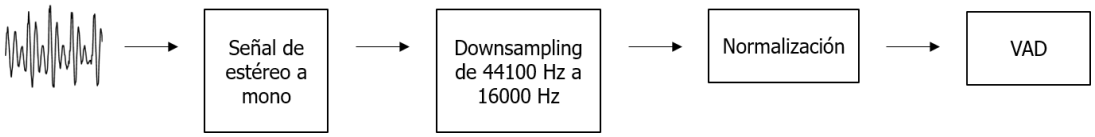

FIGURA 4.2: Fase de preprocesado de las señales de voz

En primer lugar, se decide pasar las grabaciones de audio de estéreo a mono con el fin de facilitar el manejo de una única señal, es decir, un único vector de muestras.

Posteriormente se reduce la frecuencia de muestreo, pasando de 44100 Hz a 16000 Hz. De esta forma, disminuimos la carga computacional de nuestro sistema sin perder la calidad de la señal, que, al tratarse de voz, la elección de quedarnos con 16000 muestras por segundo resulta suficiente.

El siguiente paso es el que permite comparar unas señales con otras, es decir, la normalización. En esta fase, se busca que todas las señales tengan su amplitud en un rango de [-1,1] con media 0. Para ello, a la señal se le resta su media y, posteriormente, se divide entre el valor absoluto de su valor máximo de amplitud. En el eje y de la Figura 4.3 se puede comprobar el proceso de normalización.

El último paso es posiblemente el más importante del preprocesado de las señales. En esta fase se diseña un *Voice Activity Detector (VAD)* con el objetivo de detectar las zonas de silencio de la señal y eliminarlas, ya que estas zonas no contienen características útiles para nuestro proyecto.

El diseño del VAD se realizó con la función *vadhson*, perteneciente a la toolbox Voicebox. La misión de esta función es, a partir de una señal de audio, obtener un vector de probabilidades en un rango de 0(silencio) a 1(sonido) para cada muestra de la señal. Tras obtener este vector de probabilidades, se optó por hacer un procesado del mismo antes de usarlo directamente para recortar la señal. Esto se decidió así ya que, para no perder el sincronismo entre muestras de audio y muestras de HR, y teniendo en cuenta que los valores de Zecg se presentaban cada 1 segundo, los recortes de silencio de los audios debían realizarse también en bloques de 1 segundo.

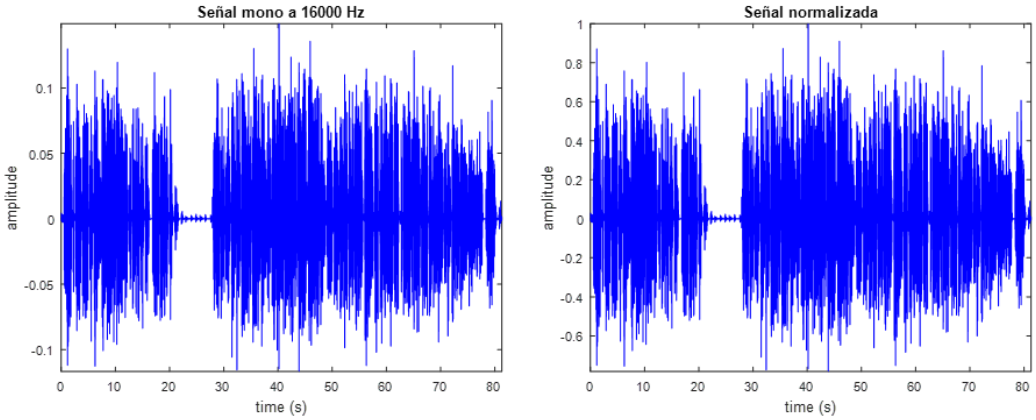

FIGURA 4.3: Representación, mediante la herramienta MIRToolbox, del proceso de normalización para la señal del instante *recording* del ID 62963719

Por tanto, a partir del vector de probabilidades se calcularon dos nuevos vectores binarizados, es decir, con valores 1 (sonido) o 0 (silencio): uno de ellos para recortar los silencios de la señal de audio, y otro para eliminar los silencios de los correspondientes vectores de HR, como se detallará en la Sección 4.4. En la Figura 4.4 se muestran los pasos que se han seguido para calcular estos vectores, a partir del obtenido de la función de Voicebox, así como una representación visual del formato de cada uno de ellos.

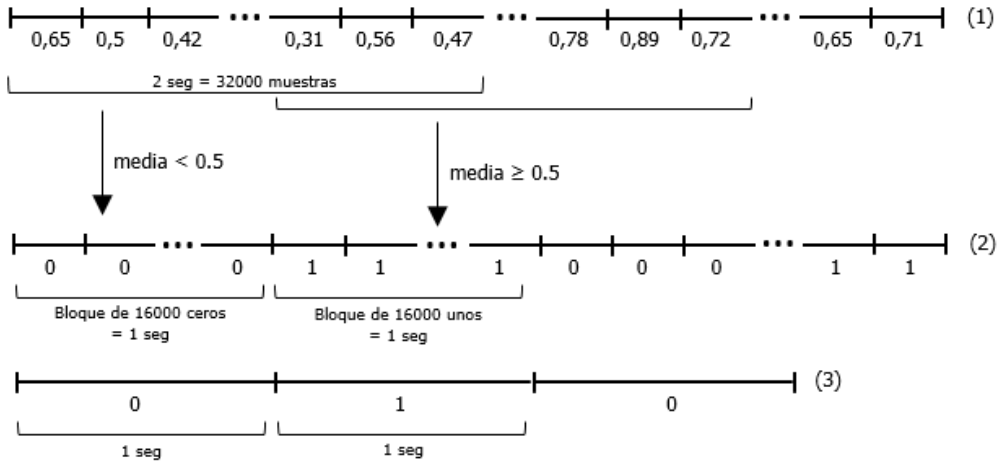

FIGURA 4.4: Representación del (1) vector de probabilidades obtenido de la función de Voicebox, (2) el vector de sonido/silencio para la señal de audio y (3) el vector de sonido/silencio para valores de HR

En el caso de la señal de voz, eliminamos los silencios de archivo quedándonos con aquellas muestras que contienen sonido, es decir, las muestras cuya posición correspondiente en el vector de sonido/silencio sea 1. En la Figura 4.5 se muestra un ejemplo del recorte de las zonas de silencios.

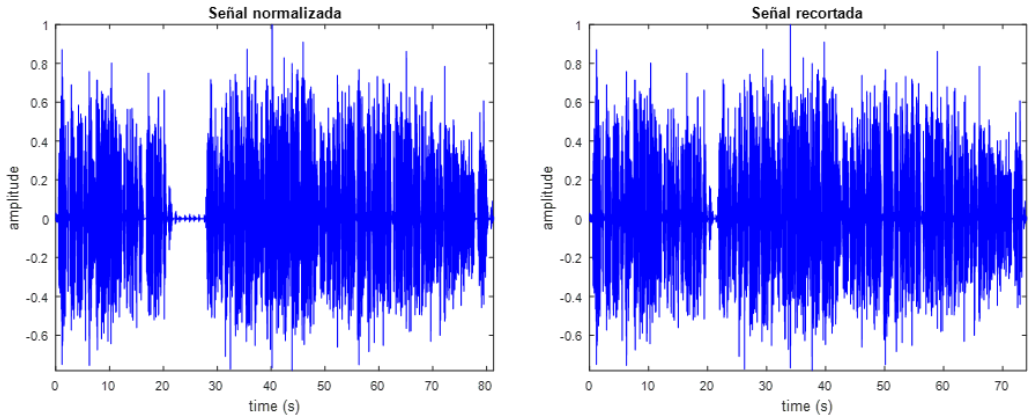

FIGURA 4.5: Ejemplo de eliminación de las zonas de silencios para la señal del instante *recording* del ID 62963719

Tras esta fase de preprocesado, las señales ya estan listas para el paso de extracción de características, detallado en la Sección 4.5.

4.4. Generación de etiquetas

En la mayoría de sistemas, la fase de generación de etiquetas es un proceso independiente del preprocesado de la señal, es decir, podría realizarse en paralelo o incluso no tener que hacerlo, ya que muchas bases de datos cuentan con las etiquetas por separado. Sin embargo, en nuestro caso, no solo tenemos que realizar esta fase sino que depende de la anterior ya que, como recordamos, hemos obtenido un vector de sonido/silencio con la dimensión correspondiente para poder eliminar las zonas de silencio del vector de HR.

Por tanto, en este punto del proyecto disponemos de la siguiente información para cada ID: 2 señales de voz listas para la extracción de características, 2 vectores con valores de Zecg o, lo que es lo mismo, HR, para los instantes *base* y *recording* y, por último, 2 vectores sonido/silencio, que contienen 1s y 0s, correspondientes a los anteriores vectores de HR.

Pero antes de pasar al cálculo de las etiquetas, en este punto del proceso es importante recordar los dos niveles de análisis que planteábamos en el Capítulo 3, es decir, los dos tamaños de segmento con los que se analizaría el sistema. Esta cuestión afecta tanto a la generación de etiquetas como a la extracción de características, que se detallará más adelante en la Sección 4.5.

En el proyecto se propone trabajar los siguientes niveles de análisis:

- **Nivel 1:** Tamaño de ventana de 2 segundos con desplazamiento de 1 segundo.
- **Nivel 2:** Aumento del nivel 1 en un factor de 5, es decir, tamaño de ventana de 10 segundos con desplazamiento de 5 segundos.

El objetivo de realizar estos dos niveles de segmento es comprobar si los resultados de las pruebas que más tarde realizaremos son mejores o peores según la resolución de las etiquetas y características.

En cuanto a la generación de etiquetas, el primer paso es el de establecer el/los umbral/es de decisión, es decir, elegir cómo vamos a determinar cuándo, en una muestra de audio, hay estrés o no.

En nuestro estudio se proponen dos umbrales diferentes, únicos para cada ID, ambos basados siempre en los valores de HR del archivo *base*, ya que es el que tomamos como situación de reposo del hablante. El primero que proponemos, que denominaremos *umbral 1* se basará en el valor resultante de sumar la media y la desviación típica de los valores de HR del ID correspondiente. Por otro lado, el segundo, llamado *umbral 2* se inspira en la idea utilizada por Mariana Dimas [21] de utilizar un percentil, es decir, a partir de un porcentaje fijado, en nuestro caso establecemos el 75 %, se buscarán las muestras de HR con valores por encima o por debajo de ese percentil.

Por tanto, aplicamos el criterio de que, si el valor de HR de una muestra supera el umbral, la etiqueta correspondiente será un 1 (estrés) y por el contrario, si el valor está por debajo del umbral, la etiqueta será un 0 (no estrés). De esta manera, obtendremos para el Nivel 1 los vectores *labels11* y *labels12*, descritos en la Sección 3.2, para cada uno de los instantes *base* y *recording*, ambos con una resolución de 1 etiqueta/segundo.

Tras calcular los vectores de etiquetas basándonos en los vectores de HR y los umbrales fijados, pasamos al último paso: recortar las zonas de silencio de dichos vectores para que se correspondan de manera sincronizada con las señales de voz. Este recorte lo realizamos de la misma manera que hicimos con las señales de voz ya que las dimensiones encajan debido a que la resolución en este caso es la misma, es decir, 1 muestra por segundo. Nos quedaremos con aquellas muestras cuya posición correspondiente en el vector de sonido/silencio sea 1. En la Figura 4.6 se representa este proceso de recorte.

Por último, queda por realizar el cálculo de etiquetas para el nivel 2. Estas etiquetas, como se explicó anteriormente, se calcularán a partir del nivel 1. En este caso, el criterio para calcularlas será, aplicando la ventana de nivel 2, que la etiqueta resultante sea un 0 o un 1 por mayoría según el número de 1s o 0s que contenga la ventana, dando prioridad al 1 (estado estrés) en los casos en los que el número de 1s sea igual al de 0s. En la Figura 4.7 se muestra dicho cálculo. Con este proceso obtenemos, para el nivel 2, los vectores *labels21* y *labels22*, para cada uno de los instantes *base* y *recording*.

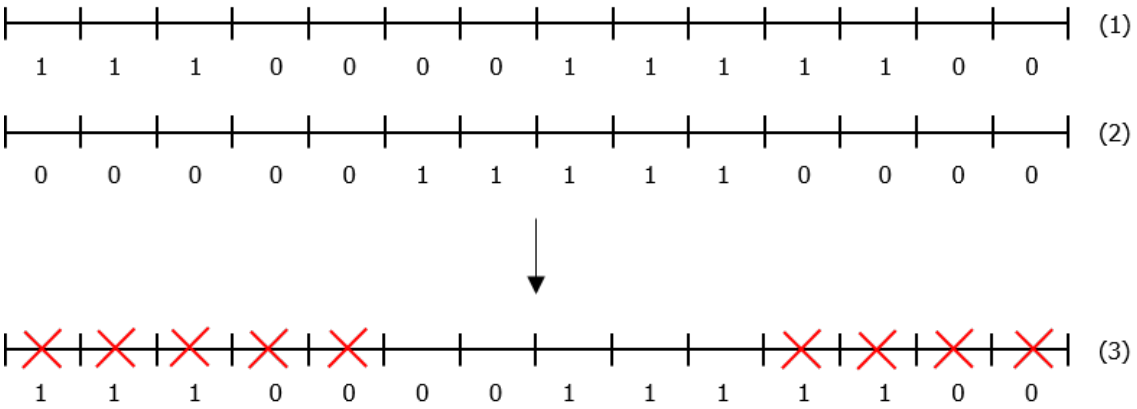

FIGURA 4.6: Representación de (1) el vector de etiquetas original, (2) el vector de sonido/silencio para los valores de HR y (3) el vector de etiquetas con una X en las zonas de silencio que se van a eliminar, correspondientes a las posiciones de 0s del (2).

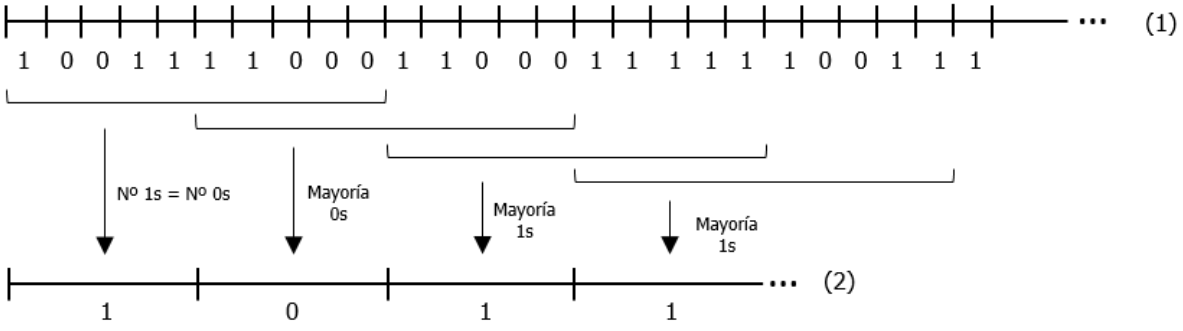

FIGURA 4.7: Representación del cálculo de etiquetas de Nivel 2 (2) a partir de las de Nivel 1 (1), mediante mayoría, dando prioridad al 1 (estrés)

4.5. Extracción de características

Como se ha descrito en capítulos anteriores, la fase de extracción de características es la pieza clave de todo el sistema ya que, gracias a ella, obtendremos la información que actuará como entrada de nuestro clasificador. Esta etapa, al igual que las anteriores, se ha realizado con la herramienta MATLAB.

Debido a problemas de memoria con el programa, los audios con duración mayor de 4 minutos se dividirán en bloques de 4 minutos con el fin de evitar problemas de computación en esta fase de extracción.

Antes de comenzar con las características, recordamos que, al igual que en la fase de generación de etiquetas, aquí también aparecen los mismos dos niveles de ventana o segmento. De esta manera, las dimensiones tanto de etiquetas como de las características coincidirán, siendo esto un requisito indispensable para la etapa de clasificación.

Basándonos en los estudios de la Sección 2.6, nuestro proyecto propone la extracción de un set de características básicas así como de sus estadísticos con el fin de obtener la información suficiente para la detección de estrés. A continuación se detalla el proceso de extracción de cada una de ellas:

- **Pitch.** El pitch, también conocido como frecuencia fundamental f_o, es una característica que se puede extraer mediante diferentes métodos: autocorrelación, análisis del espectro, método SIFT, etc. En nuestro caso, extraeremos el pitch utilizando la función *fxpefac* de la herramienta VoiceBox. Esta función se basa en el cálculo del pitch mediante la técnica Pitch Estimation Filter with Amplitude Compression (PEFAC) [43], un algoritmo robusto indicado para altos niveles de ruido. Gracias a la salida de esta función podemos discriminar posteriormente en el vector de pitch las zonas sonoras de las sordas. Esto resulta de gran utilidad ya que de estas últimas no es interesante el cálculo de f_o. Por tanto, el resultado que obtenemos es un vector de pitch, en unidades de Hercios, donde la frecuencia será 0 para las zonas sordas de la señal.

- **MFCC.** Los coeficientes MFCC se han calculado, de la misma manera que el vector de pitch, mediante la herramienta VoiceBox, en este caso con la función *melcepst*. Como resultado de esta función, conseguimos una matriz de 12 filas, cada una de ellas asociada a un coeficiente MFCC.

- **Frecuencias formantes.** Los formantes, a diferencia de las características anteriores, se han extraído ad hoc basándonos en la teoría de LPC. La predicción lineal modela la señal como si fuera generada por una señal de energía mínima que pasa a través de un filtro IIR puramente recursivo, obteniendo de esta manera la respuesta en frecuencia de la Figura 2.2. Para encontrar las frecuencias formantes del filtro, necesitamos encontrar las localizaciones de las resonancias del mismo. Esto implica tratar los coeficientes del filtro como un polinomio y resolver sus raíces, obteniendo así las frecuencias formantes. En nuestro caso, nos quedaremos únicamente con los tres primeros formantes ya que son los que proporcionan más información. De esta manera, obtenemos una matriz de 3 filas, cada una de ellas correspondientes a cada uno de las frecuencias formantes.

Para la extracción de las tres características anteriores se aplicó una ventana de análisis de 20 ms con desplazamiento de 10 ms, aproximadamente como se suele calcular de forma clásica.

Sin embargo, las dimensiones de los vectores de características obtenidos con esta ventana de análisis no se corresponden con las que necesitamos para el nivel 1 (ventana de 2 segundos con desplazamiento de 1 segundo). Para adaptar las dimensiones de estos vectores, se optó por calcular estadísticos como la media y la varianza para cada vector de características, aplicando, en este caso, la ventana correspondiente al nivel 1. De esta manera, logramos las dimensiones correctas y, además, ampliamos el conjunto de características para la posterior clasificación. En la Figura 4.8 representamos este proceso de redimensión.

- **Energía.** Por último, calculamos, para el nivel 1, la energía de la señal como el módulo al cuadrado de la misma en el intervalo de muestras que contenga la ventana:

$$E = \sum_{n=-\infty}^{\infty} |x(n)|^2$$

Tras calcular todos los vectores de características y apilarlos unos tras otros, obtenemos una matriz *features1* de dimensiones 33xN1, donde N1 es el número de tramas correspondientes al primer nivel de ventana. En la Tabla 4.3 se desglosa, por filas, la matriz de características a nivel 1.

Fila	Caracerística
1	Vector medias pitch (1xN1)
2	Vector varianzas pitch (1xN1)
3-14	Matriz medias MFCC (12xN1)
15-26	Matriz varianzas MFCC (12xN1)
27-29	Matriz medias formantes (3xN1)
30-32	Matriz varianzas formantes (3xN1)
33	Vector energía (1xN1)

CUADRO 4.3: Desglose, por filas, de la matriz de características a Nivel 1

Por último, se calcula la matriz de características a nivel 2, *features2*. Para ello, calcularemos, de forma similar al proceso de generación de etiquetas a nivel 2, los vectores de características basándonos en el nivel 1. A diferencia de la fase anterior, la ventana no evaluará la mayoría de unos o ceros, sino que calculará la media aritmética de los valores que contenga a nivel 1. De esta manera, obtenemos una matriz de dimensiones 33xN2, donde N2 es el número de tramas correspondientes a nivel 2.

En la Figura 4.8 se representa el proceso completo de extracción de características para un vector completo, es decir, el paso de la ventana de 20 ms, a la de nivel 1, y posteriormente a la de nivel 2.

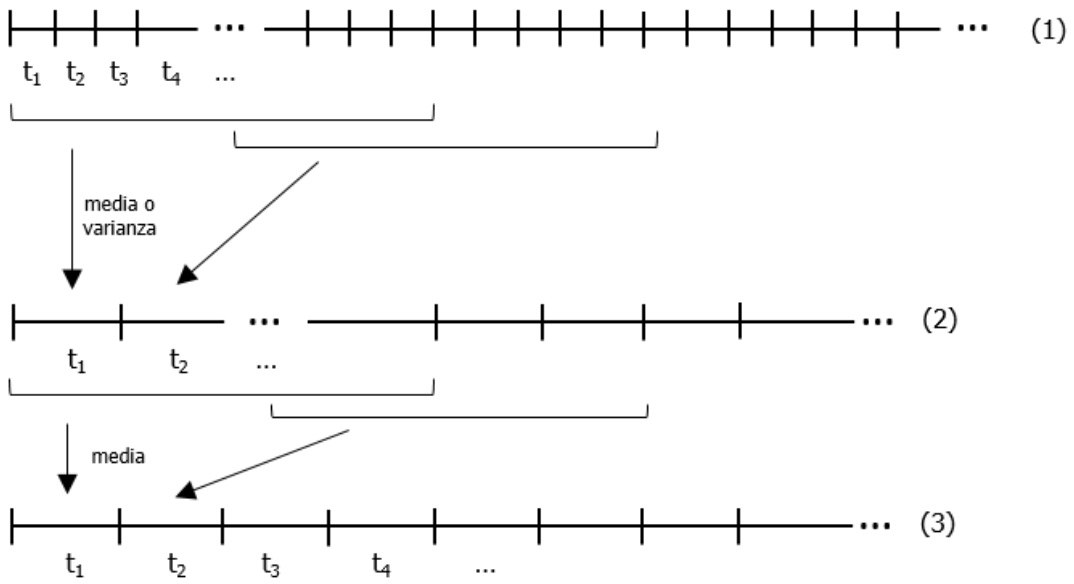

FIGURA 4.8: Representación de (1) un vector de características para una ventana de 20 ms, (2) el vector de medias o varianzas para el nivel 1 y (3) el vector de medias o varianzas para el nivel 2; donde t_i representa cada trama del vector

4.6. Clasificación y evaluación

En esta última etapa utilizaremos toda la información generada con anterioridad como entrada a nuestro clasificador. Tras esta fase, podremos evaluar los resultados obtenidos a partir de distintas medidas de rendimiento y, así, determinar qué sistema funciona mejor.

Antes de ello, recordamos los datos con los que contamos en este punto del proyecto. Disponemos de 2 sets de datos, Set 1 y Set 2, ambos con diferentes IDs, es decir, participantes. Para cada set, disponemos de la siguiente información:

- *features1:* Matriz de características extraídas de todos los archivos de audio, tanto del instante *recording* como *base*, para el nivel 1.
- *labels11:* Vector de etiquetas de todos los archivos de HR calculadas sobre el umbral 1, correspondientes a cada una de las tramas de la matriz de características, a nivel 1.
- *labels12:* Vector de etiquetas de todos los archivos de HR calculadas sobre el umbral 2, correspondientes a cada una de las tramas de la matriz de características, a nivel 1.
- *features2:* Matriz de características extraídas de todos los archivos de audio, tanto del instante *recording* como *base*, para el nivel 2.

- *labels21:* Vector de etiquetas de todos los archivos de HR calculadas sobre el umbral 1, correspondientes a cada una de las tramas de la matriz de características, a nivel 2.
- *labels22:* Vector de etiquetas de todos los archivos de HR calculadas sobre el umbral 2, correspondientes a cada una de las tramas de la matriz de características, a nivel 2.

Como para diseñar un clasificador necesitamos una matriz de características y un vector de etiquetas, se realizarán, para cada máquina, las 4 posibles combinaciones entre características y etiquetas, 2 en cada nivel.

A continuación se detallará el proceso de la fase de clasificación para una de las combinaciones, siendo igual para las tres restantes. Esta parte del proyecto se ha desarrollado con el lenguaje de programación Python.

El primer paso, como en todo sistema de clasificación, es dividir los datos de forma completamente aleatoria en dos conjuntos: train (80 %) y test (20 %).

Tras esto, para que durante la fase de entrenamiento tanto la clase 1 (estrés) como la 0 (no estrés) tengan el mismo peso, es decir, la misma importancia, se realiza el equilibrado del conjunto de entrenamiento. La forma en la que se ha decidido hacer esta fase es mediante clonación, es decir, una vez detectada la clase de la que hay menos muestras, se toman algunas de ellas de forma aleatoria, se clonan, y se añaden al conjunto de nuevo.

Posteriormente, se realiza la normalización de los conjuntos de entrenamiento y de test. Para ello, a ambos se les resta la media del set de train y se dividen entre la desviación típica del mismo.

Para el diseño de las máquinas, tanto para MLP como SVM, decidimos utilizar los valores por defecto del paquete *sklearn* de Python, es decir, valores de penalización de $alpha = 10^{-4}$ para el caso de MLP y de $C = 1$ en SVM. Hasta el momento no hemos hecho ningún intento por adaptar estos valores de configuración y este es el motivo por el que no hemos separado un conjunto de validación. En el Capítulo 5 se darán más detalles de las pruebas realizadas con cada máquina.

Una vez entrenamos las máquinas con los datos de train, hacemos la clasificación, es decir, predecimos las etiquetas para el conjunto de test.

Para evaluar el rendimiento de nuestra máquina, o lo que es lo mismo, averiguar cómo de bien hemos realizado esta predicción, compararemos las etiquetas estimadas con las originales de test. Sin embargo, debido a que el conjunto de test no está equilibrado, es decir, no contiene el mismo número de muestras de la clase 0 que de la clase 1, no utilizaremos medidas clásicas como la *mean accuracy*, sino métricas de las curvas *precision-recall*.

En la Figura 4.9 observamos que tanto los grupos A como D se clasifican correctamente, sin embargo, no ocurre lo mismo con los casos B como C. Las métricas que utilizaremos para la evaluación se basan en estos conjuntos. Estas medidas son:

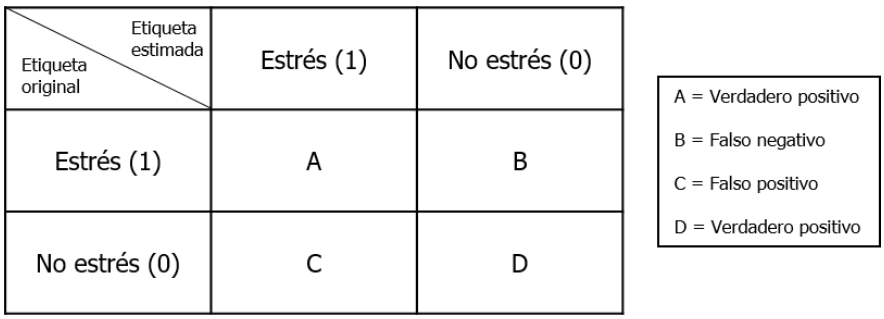

FIGURA 4.9: Representación de los grupos de clasificación en un caso binario

- Precision (P) se define como el número de verdaderos positivos (T_p) entre la suma de los verdaderos positivos (T_p) y los falsos positivos (F_p):

$$P = \frac{T_p}{T_p + F_p}$$

- Recall (R) se define como el número de verdaderos positivos (T_p) entre la suma de los verdaderos positivos (T_p) y los falsos negativos (F_n):

$$R = \frac{T_p}{T_p + F_n}$$

- F Score (F) se define como la media armónica entre Precision y Recall:

$$F = 2 \times \frac{P \times R}{P + R}$$

El criterio con el que evaluaremos nuestro clasificador será, principalmente, con el valor de F Score, de forma que cuanto más se acerque a 1, mejor habrá sido nuestra clasificación.

Capítulo 5

Pruebas y Resultados

En este capítulo se detallarán las pruebas y los resultados correspondientes que se han realizado durante el proyecto para evaluar la calidad del detector de estrés.

En primer lugar, se ha decidido dividir las pruebas en dos grandes grupos: *speaker dependent* (dependiente del locutor) y *speaker independent* (independiente del locutor). Esto hace referencia a los conjuntos que se eligen para el entrenamiento y el test del clasificador, es decir, si, al realizar la clasificación, la máquina ya conoce a los participantes de la fase de entrenamiento o si son muestras completamente desconocidas para ella. Las pruebas se realizarán combinando los distintos clasificadores, con los 2 niveles de segmento y los dos umbrales de clasificación detallados en el Capítulo 4. Como comprobaremos más adelante, esto influirá en los resultados. Cada una de estas pruebas se ha realizado 5 veces con el fin de obtener un valor medio de clasificación.

Por otro lado, se incluirán unas gráficas comparativas que representarán los datos de clasificación para cada participante o ID individual, junto con sus datos de ritmo cardíaco con el fin de encontrar una relación entre ellos.

Como se explicó en el Capítulo 4, el criterio con el que vamos a evaluar la clasificación es mediante los valores Precision, Recall y F Score. Sin embargo, a continuación se mostrarán únicamente los valores de este último, ya que se trata de una media armónica entre Precision y Recall, dejando los otros dos valores restantes como información adicional en el Apéndice A.

5.1. Pruebas dependientes del locutor

Las pruebas dependientes del locutor se basan, esencialmente, en utilizar el mismo conjunto de participantes o IDs, tanto para la fase de entrenamiento como para la fase de clasificación. Para ello, las muestras se separan de forma aleatoria en 80 % train y 20 % test.

En la Tabla 5.1 se muestran los resultados F Score de las pruebas para el Set 1 de datos. Se espera que estas muestras, elegidas por tener los audios y datos de mayor calidad, den los mejores resultados de clasificación del sistema.

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.902	0.906	0.932	0.908
SVM linear	0.810	0.800	0.887	0.872
SVM poly	0.874	0.889	0.942	0.901
SVM rbf	0.908	0.910	0.942	0.925
SVM sigmoid	0.702	0.699	0.787	0.757

CUADRO 5.1: F Score prueba dependiente del locutor Set 1

A continuación, la Tabla 5.2 recoge los resultados F Score de las pruebas para el Set 2 de datos. Intuitivamente, debido a que la calidad de las muestras es peor en este conjunto, sería de esperar que los resultados fueran inferiores.

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.860	0.869	0.884	0.899
SVM linear	0.744	0.770	0.789	0.809
SVM poly	0.856	0.858	0.863	0.870
SVM rbf	0.855	0.860	0.860	0.880
SVM sigmoid	0.626	0.632	0.595	0.619

CUADRO 5.2: F Score prueba dependiente del locutor Set 2

La Tabla 5.3 contiene los resultados para el Set 3, es decir, para la suma de muestras del Set 1 y el Set 2. Estos resultados resultan interesantes de comparar con la Tabla 5.4 de la Sección 5.2 para evaluar, en lugar de cómo afecta la calidad de los audios como pasaba en las pruebas anteriores, el grado de dependencia del hablante que existe en nuestro sistema. Esto lo comprobaremos según la variabilidad de los resultados entre una prueba y otra.

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.862	0.864	0.906	0.911
SVM linear	0.729	0.744	0.787	0.831
SVM poly	0.856	0.862	0.897	0.900
SVM rbf	0.860	0.863	0.893	0.900
SVM sigmoid	0.592	0.591	0.591	0.643

CUADRO 5.3: F Score prueba dependiente del locutor Set 3

A la vista de los resultados obtenidos en las pruebas dependientes del locutor, observamos que, para las pruebas del Set 1 y Set 2, obtenemos los mejores resultados con los clasificadores SVM con función Kernel RBF. Sin embargo, las pruebas con el Set 3, en las que el número de muestras crece, obtienen mejores resultados con un clasificador MLP.

En cuanto a los diferentes niveles y umbrales, aparentemente los resultados a nivel 2, es decir, cuando la ventana de análisis es mayor, son ligeramente superiores al resto. En el caso de los umbrales de decisión, parece destacar el umbral 1 para el Set 1, donde la calidad de los datos es mayor, y el umbral 2 para el Set 2 y 3 que contienen información de peor calidad.

5.2. Pruebas independientes del locutor

En nuestro proyecto, las pruebas independientes del locutor consisten en diseñar un sistema de detección entrenado con un conjunto de IDs para, posteriormente, realizar el test con otro conjunto de participantes totalmente desconocidos para la máquina.

En la Tabla 5.4 se recogen los resultados para una máquina cuyo entrenamiento se ha realizado con los datos del Set 1 y la etapa de test con el Set 2.

	Nivel 1 umbral 1	**Nivel 1 umbral 2**	**Nivel 2 umbral 1**	**Nivel 2 umbral 2**
MLP	0.820	0.839	0.831	0.850
SVM linear	0.666	0.677	0.742	0.770
SVM poly	0.774	0.793	0.819	0.841
SVM rbf	0.810	0.833	0.822	0.836
SVM sigmoid	0.668	0.669	0.649	0.708

CUADRO 5.4: F Score prueba independiente del locutor entre Set 1 y Set 2

Los resultados de la Tabla 5.4 reflejan que el clasificadores que proporcionan una detección mejor es el MLP. En este caso, al igual que ocurría en las pruebas dependientes del locutor donde la calidad de los datos es peor, la combinación de nivel 2 y umbral 2 es la que proporciona una detección más fiable.

Si comparamos la Tabla 5.4 con la Tabla 5.3, observamos que esta primera recoge resultados ligeramente más bajos respecto a las dependientes del locutor. Esto significa que en nuestro sistema sí existe cierta dependencia con el locutor. Sin embargo, la diferencia no es demasiado drástica como para que esto se convierta en un factor determinante.

5.3. Gráficas comparativas

En las pruebas anteriores podíamos extraer una relación entre la calidad de los diferentes sets y la forma en la que se realizaba la clasificación, bien por el nivel, el umbral o el tipos de clasificador. Sin embargo, otro de los objetivos interesantes de nuestro estudio era el de buscar un vínculo entre los valores de ritmo cardíaco y los de la detección o F Score.

Para ello, se analizaron distintas formas de representar los datos. A la izquierda de la Figura 5.1 podemos observar, para cada ID individual representado en el eje X[1] , los valores medios de HR tanto en el instante *base* como en *recording* así como sus correspondientes valores de F Score. Tras analizar este tipo de gráfica, se decidió diseñar una nueva que representase, a diferencia de la gráfica anterior, la resta de valores de HR del instante *recording* menos los del instante *base*; así como el F score, como vemos a la derecha de la Figura 5.1. De esta forma, se pueden sacar conclusiones con mayor claridad.

[1]Por simplificar la figura se ha decidido numerar los IDs a partir del número 1. En el Apéndice A podemos encontrar la correspondencia de cada número del eje X con su respectivo ID según el set de datos.

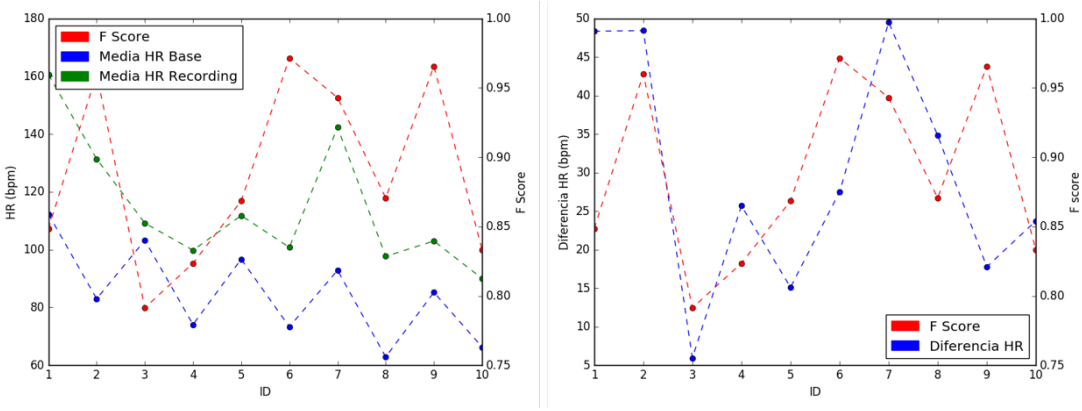

FIGURA 5.1: Gráficas comparativas para Set 1, clasificador MLP, Nivel 1 Umbral 1, sin tener en cuenta el género

En la Figura 5.2 representamos esta prueba para el nivel 1, es decir, con una ventana de análisis pequeña. Cabe señalar que en la gráfica del set 3 aparecen en primer lugar los 10 IDs pertenecientes al set 1 (del 1 al 10) y a continuación los 14 pertenecientes al set 2 (del 11 al 24).

En esta figura podemos observar varias cosas. Por una parte, los IDs que obtienen mala clasificación para el set 2, tienden a mejorar cuando se combinan con las muestras del set 1, es decir, cuando se clasifican en el set 3. Esto podría deberse a que un aumento en el número de muestras, que además son de mayor calidad, proporcionaría un mejor diseño de la máquina y, por tanto, una mejor clasificación. Sin embargo, la detección de los IDs del set 1 no varía de forma destacable cuando los analizamos en el set 3.

Por otro lado, de forma general, observamos en este nivel 1 aparece cierta relación entre los valores de F score y la diferencia de HR, de forma que, excepto casos aislados, parecen directamente proporcionales ya que para diferencias de HR bajas aparecen F scores menores.

Al contrario que en la figura anterior, en la Figura 5.3 mostramos los datos de los 3 sets para el nivel 2, es decir, la tamaño de segmento mayor.

En esta Figura 5.3 encontramos varias diferencias respecto a la anterior Figura 5.2. La primera de ellas, como habíamos comprobado en la Sección 5.1, los valores de F Score resultan, de forma general, mayores en el nivel 2 que en el nivel 1. Otra diferencia significativa que observamos es que en el nivel 2 no parece existir la relación que sí había a nivel 1 entre la diferencia de HR y el F Score.

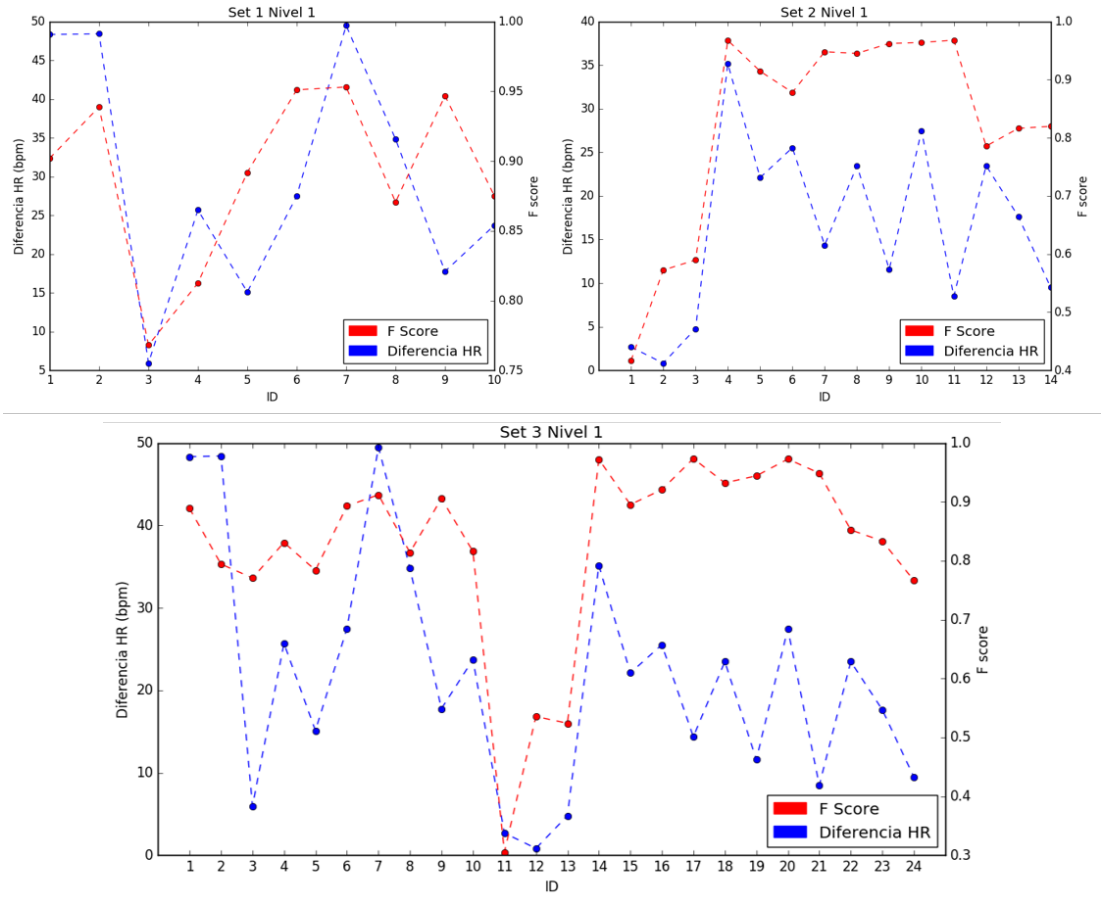

FIGURA 5.2: Gráficas comparativas para Set 1, 2 y 3, clasificador MLP, Nivel 1 Umbral 1, sin tener en cuenta el género

Tanto en la Figura 5.2 como en la Figura 5.3 observamos un ID concreto que recoge valores de clasificación destacablemente bajos. Este ID se corresponde con el 12782919 y, tras analizar sus valores de HR de forma manual, se ha observado que, tanto en el instante *base* como *recording*, los valores de HR, que son prácticamente iguales, son relativamente bajos comparados con el resto de participantes. Por tanto, esta puede ser la razón por la que su clasificación es peor, ya que es la única muestra con estas características.

En cuanto a las pruebas entre clasificadores, no se han encontrado diferencias destacables entre las gráficas del MLP o SVM con función de kernel RBF.

Las gráficas anteriores se han realizado para el umbral de decisión 1. Tras realizar estas mismas pruebas para el umbral 2, el único aspecto a destacar es que para este último los valores de las gráficas, entre los dos niveles, suelen variar de forma más brusca respecto al umbral 1.

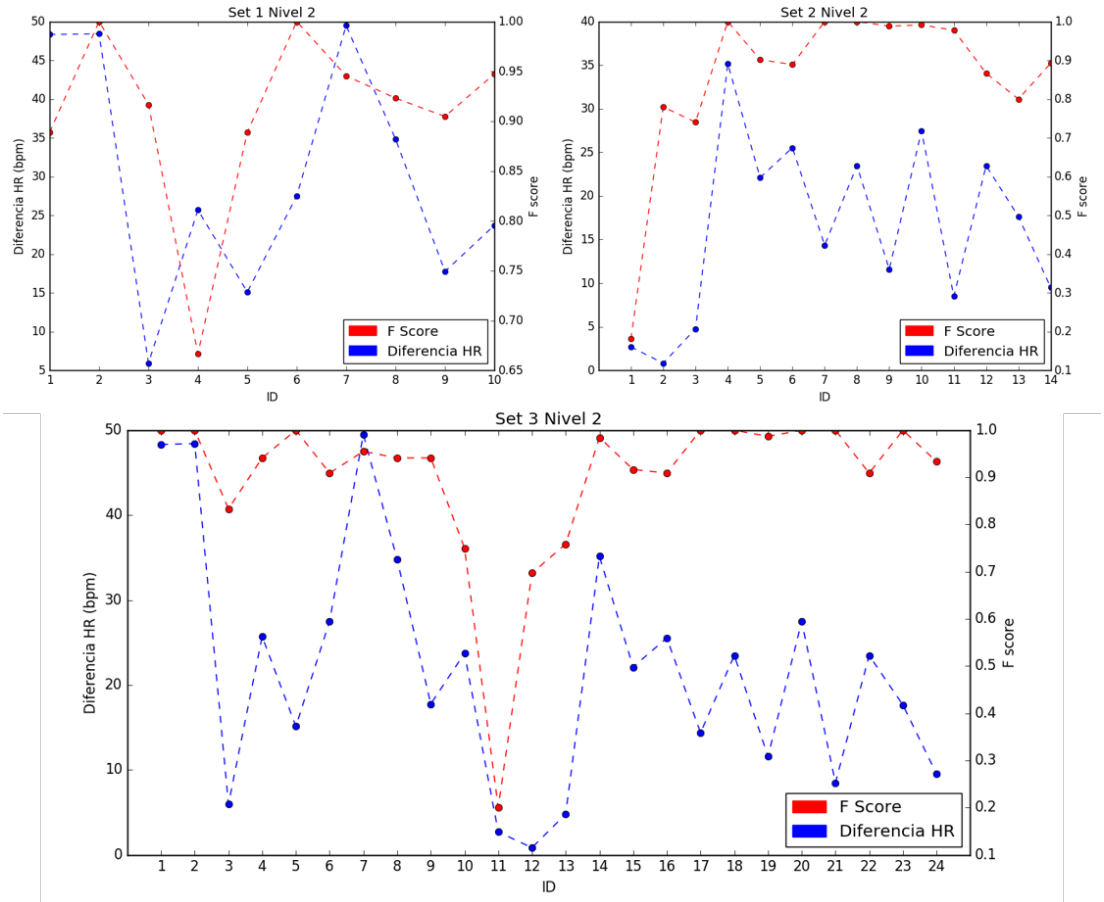

FIGURA 5.3: Gráficas comparativas para Set 1, 2 y 3, clasificador MLP, Nivel 2 Umbral 1, sin tener en cuenta el género

Por último, otro objetivo de este proyecto era el de responder a la pregunta de si existen diferencias entre hombres, mujeres y edades a la hora de realizar la detección.

En cuanto a las edades, debido a que los participantes se encontraban en el mismo rango de edad, no se han obtenido diferencias concluyentes en las pruebas.

En cuanto al género, la limitación de información de nuestra base de datos, de la que solo conocíamos el género de 38 de los 45 participantes, y a los IDs que posteriormente hemos descartado, ha supuesto que no se han encontrado diferencias significativas entre la detección en hombres y en mujeres. En las Figura 5.4 y Figura 5.5 mostramos las gráficas correspondientes a los 14 hombres y 7 mujeres del set 3.

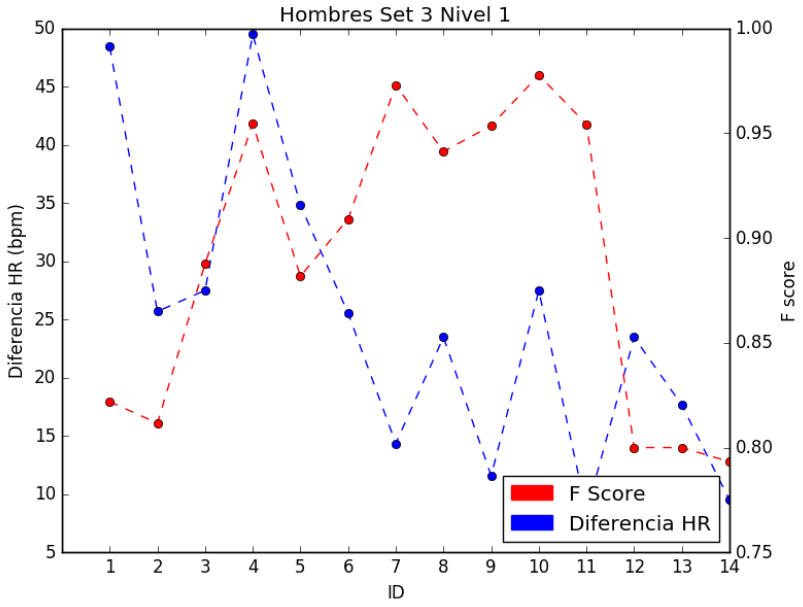

FIGURA 5.4: Gráficas comparativas para Set 3, clasificador MLP, Nivel 1 Umbral 1, género masculino

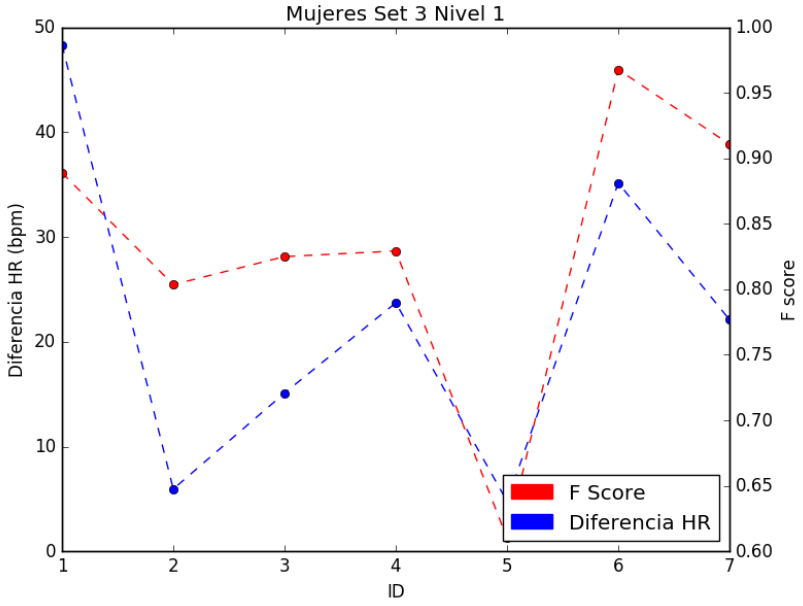

FIGURA 5.5: Gráficas comparativas para Set 3, clasificador MLP, Nivel 1 Umbral 1, género femenino

Capítulo 6

Planificación y presupuesto

6.1. Planificación

Debido al carácter vital que tiene la etapa de planificación para cualquier proyecto con éxtio, el objetivo de esta sección será detallar la la organización, tareas y ejecución temporal de nuestro estudio.

Para ello, se recogen en la Figura 6.1 el desglose de las tareas que han compuesto el proyecto así como su fecha de comienzo, de fin y su duración en días.

Para facilitar la comprensión del proceso, se va a acompañar esta información con el diagrama de Gantt de la Figura 6.2 que indicará de manera gráfica la duración de cada tarea y la secuencia seguida, así como el tiempo que ha sido dedicado a cada tarea en comparación con el tiempo total del proyecto.

	Nombre de tarea	Duración	Comienzo	Fin
1	Obtención base de datos	6 días	vie 10/02/17	vie 17/02/17
2	Procesado de la base de datos	23 días	sáb 18/02/17	mar 21/03/17
3	Creación de sets de datos	4 días	mié 22/03/17	lun 27/03/17
4	Decisión de umbrales y niveles	5 días	lun 27/03/17	vie 31/03/17
5	Prerocesado de las señales de voz	16 días	sáb 01/04/17	vie 21/04/17
6	Generación de etiquetas	16 días	mar 04/04/17	mar 25/04/17
7	Extracción de características	15 días	jue 13/04/17	mié 03/05/17
8	Diseño de clasificadores	4 días	jue 04/05/17	mar 09/05/17
9	Redacción memoria	27 días	mié 10/05/17	jue 15/06/17
10	Pruebas clasificación	5 días	vie 02/06/17	jue 08/06/17

FIGURA 6.1: Listado de tareas del proyecto

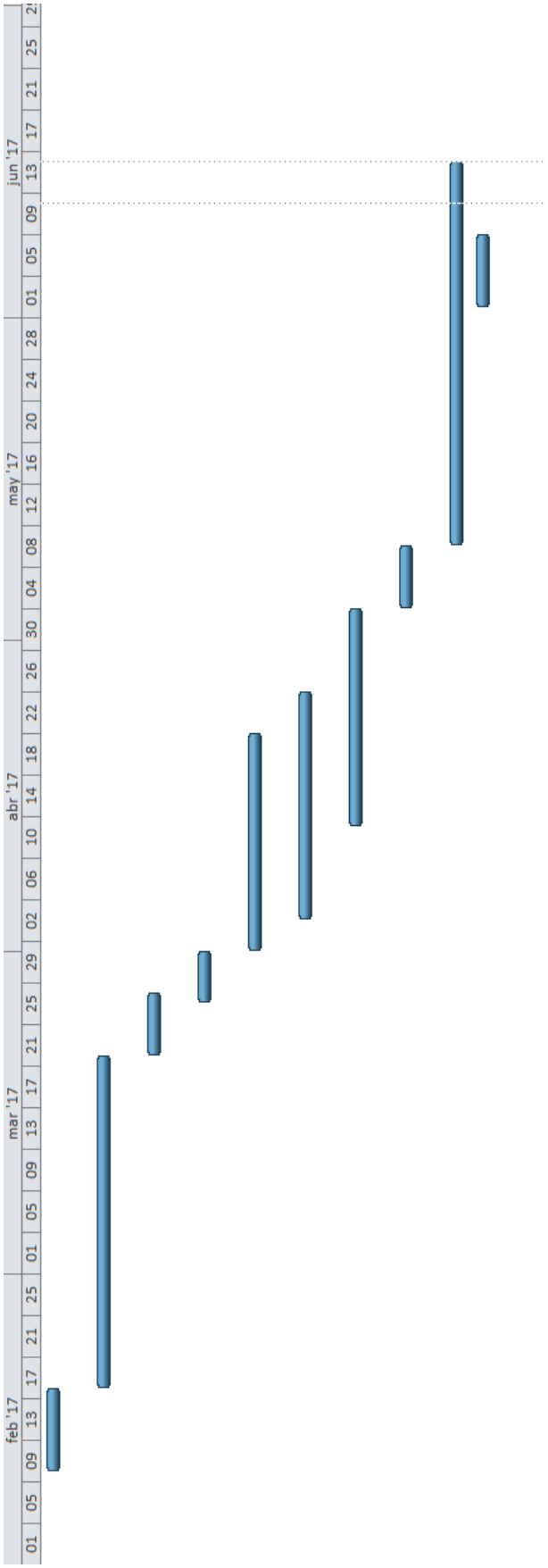

FIGURA 6.2: Tareas del proyecto representadas en un diagrama de Gantt

6.2. Presupuesto

En esta sección se hará una estimación del coste del proyecto. Para ello realizará un listado de los recursos utilizados junto con sus costes correspondientes.

- **Recursos físicos.**
  En nuestro caso, el único recurso físico utilizado ha sido un ordenador portátil Acer Aspire E1-571 cuyo coste total es de 393.39€ sin IVA. Si consideramos una amortización lineal durante 10 años, el coste del equipo para un año sería 39.34 €, como refleja la Tabla 6.1.

Recurso	Coste amortización
Ordenador portátil Acer Aspire E1-571	39.34 €

CUADRO 6.1: Tabla presupuesto recursos físicos

- **Recursos software.**
  En la Tabla 6.2 se recogen los costes asociados a los recursos software utilizados en el proyecto, es decir, sistema operativo y programas para su desarrollo.

Recurso	Coste
Microsoft Windows 10 Home 64 bits	78.47 €
MATLAB 2016 Licencia estudiante	69€
Spyder IDE - Anaconda Python	0 €
TOTAL sin IVA	**147.47 €**

CUADRO 6.2: Tabla presupuesto recursos software

- **Recursos humanos**
  En cuanto a los recursos humanos, consideraremos que en el proyecto se ha realizado por un desarrollador con nivel profesional equivalente a ingeniero junior y a un supervisor con nivel de ingeniero senior. A cada uno de ellos se le asignará un coste de 15€/hora y 20€/hora, respectivamente.

  El período de desarrollo del proyecto ha sido de aproximadamente 20 semanas, en las que se ha seguido una jornada de trabajo continua. Se ha estimado que el desarrollador ha invertido aproximadamente 18 horas/semana, mientras que el supervisor ha dedicado 2 horas/semanas para tareas de supervisión y control del proyecto. En la tabla Tabla 6.3 se reúnen los gastos totales a estos recursos humanos.

Recurso	**Horas dedicadas**	**Coste/hora**	**Coste total**
Desarrollador	18 horas/semana	15 €/hora	5400 €
Supervisor	2 horas/semana	20 €/hora	800 €
TOTAL			**6200 €**

CUADRO 6.3: Tabla presupuesto recursos humanos

En la Tabla 6.4 se muestra el resumen del presupuesto total, cuya cifra asciende a **seis mil trescientos ochenta y seis euros y ochenta y un céntimos sin IVA** .

Recurso	**Coste**
Recursos físicos	39.34 €
Recursos software	147.47 €
Recursos humanos	6200 €
TOTAL sin IVA	**6386.81 €**

CUADRO 6.4: Tabla presupuesto total

Capítulo 7

Conclusiones y líneas futuras

En este capítulo se redactan las conclusiones del proyecto relacionándolas con los objetivos propuestos en el Capítulo 1 con el fin de entender con mayor claridad la consecución de los mismos. Con estas conclusiones se presenta el final del proyecto acompañado de unas líneas futuras que indican el camino que podría seguir una continuación del mismo.

7.1. Conclusiones

Con el fin de diseñar un detector lo suficientemente preciso para eventos reales de estrés, nuestro estudio ha realizado múltiples combinaciones entre la fase de extracción de características y clasificación, buscando encontrar, de todas estas combinaciones, aquella que proporcione mejor resultado de clasificación.

Gracias a los resultados de las pruebas dependientes e independientes del locutor del Capítulo 5, podemos extraer diferentes conclusiones y relaciones entre calidad del audio, ritmo cardíaco y estrés.

Los resultados de este capítulo demuestran que la relación entre estrés, HR y voz es innegable, ya que se ha podido realizar, con un alto porcentaje de acierto, la detección de estrés en voz basando la anotación de las etiquetas en valores de ritmo cardíaco.

Igualmente, estos resultados no se habrían logrado si, además de las etiquetas, las matrices de características no hubieran sido acertadas. Podemos afirmar que el set básico de características elegidas, junto con sus estadísticos, ha sido suficiente y oportuno para realizar la detección.

Los resultados de las pruebas dependientes e independientes del locutor del Capítulo 5 demuestran, además, varios aspectos acerca de los diferentes niveles y umbrales utilizados. En cuanto a los niveles, se ha demostrado que una ventana mayor, como era de esperar, da lugar a unos resultados de detección mayores, independientemente del set y la dependencia del hablante. Sin embargo, los umbrales de detección no presentan esta uniformidad. Aunque la diferencia no es muy grande, el umbral 1 da lugar a mejores resultados para el set de datos de mayor calidad, es decir, el set 1, mientras que el umbral 2 funciona mejor en aquellos sets cuya calidad es menor, como el set 2 y set 3.

Esta calidad del audio de los sets ha afectado notablemente, de forma general en las pruebas, en la detección de estrés. Se han obtenido los mejores resultados para el Set 1, cuyos audios eran de mayor calidad, mientras que los valores más bajos se han dado para el Set 2, de peor calidad; y valores intermedios para el Set 3 que combinaba muestras de estos dos últimos.

En cuanto al diseño de máquinas, se ha comprobado que, en general, los clasificadores SVM con función de Kernel RBF dan mejores resultados para los conjuntos de muestras pequeños, al contrario que el clasificador MLP, cuyo rendimiento aumenta para sets de mayor tamaño.

La comparación entre las pruebas dependientes e independientes del locutor responde a la pregunta de qué grado de dependencia existe en la etapa de clasificación. Teniendo en cuenta que los resultados son ligeramente mayores en las pruebas dependientes, podemos afirmar que existe cierta dependencia del hablante en nuestro sistema. Sin embargo, los resultados obtenidos no son excesivamente diferentes como para que esto suponga un factor determinante en nuestro proyecto.

En cuanto a las gráficas comparativas de la sección Sección 5.3 hemos extraído diversas conclusiones. En primer lugar, cuando realizamos la detección en el nivel 1 encontramos relación entre la diferencia de ritmo cardíaco entre el instante *recording* y *base* de cada participante con su resultado de clasificación individual. Aparentemente, aquellos IDs cuya diferencia es menor, da lugar a valores más bajos de clasificación. Sin embargo, en el nivel 2 esta dependencia deja de existir.

La intención de estas gráficas era, además, la de buscar diferencias entre género y edades de los distintos participantes. Sin embargo, no se han observado evidencias notables en ninguno de los dos casos, posiblemente a causa de la limitada información y cantidad de muestras en la base de datos seleccionada.

En resumen, se concluye que, en términos generales y tras analizar la consecución de los objetivos, el proyecto ha sido realizado con éxito tras lograr, de manera precisa, la detección de estrés en señales de voz en eventos reales.

7.2. Líneas futuras

Una vez finalizado el estudio, surgen nuevas líneas de trabajo para continuar, analizar o corregir aspectos individuales del proyecto en profundidad.

Uno de los caminos clásicos sería el de ampliar la gama de características que se extraen de los audios, como por ejemplo: jitter, shimmer, velocidad de elocución, etc. La extracción de nuevas características permitiría estudiar aquellas que resultan más reveladoras para la detección de estrés. Durante nuestro proyecto se realizó un estudio preliminar de las características más influyentes en la etapa de clasificación con la herramienta RFE (Recursive Feature Elimination) del paquete sklearn de Python. Sin embargo, los resultados obtenidos no fueron concluyentes ya que el set de características era demasiado pequeño.

Con el fin de mejorar la calidad las señales de voz para la fase de extracción de características, otro de los puntos a corregir sería preprocesar aquellos audios que contienen saturación debida al micrófono. Esto influiría notablemente en la etapa de normalización de audios, en la que se divide entre el máximo de amplitud, eliminando confusiones entre un máximo de saturación y uno real.

En esta misma fase de preprocesado de los audios, otro aspecto para perfeccionar sería la eliminación de silencios en los audios, es decir, el VAD. En nuestro caso, esta eliminación se ha realizado de forma separada entre audios y etiquetas, esforzándonos en no perder el sincronismo, pero realizando un corte por bloques, en ocasiones inexacto, que puede añadir error en la detección. Para corregir esto, una de las posibles soluciones sería añadir un tercer de análisis que englobase tanto audios como vectores de etiquetas y que recortase, de forma conjunta, los silencios de ambos con mayor precisión.

Por otro lado, un aumento de muestras en la base de datos daría lugar a la posibilidad de estudiar con mayor profundidad la influencia de del género y la edad en la detección de estrés.

Con el fin de mejorar la clasificación, una tarea a realizar sería el ajuste de los parámetros de los clasificadores, ya que en el proyecto se han utilizado valores por defecto. Para realizar esto, habría que tomar un conjunto de validación o realizar una validación cruzada para cada parámetro que se quisiera ajustar. Por otra parte, en vista de la poca diferencia que existe entre los resultados del clasificador MLP y SVM con función RBF, sería interesante calcular los márgenes de confianza de cada uno de ellos con el fin extraer de forma más rigurosa qué clasificador funciona mejor en nuestro proyecto.

Por último, y con vistas a proyectos de mayor alcance, la detección de estrés basada en características de la voz y medidas biométricas podría dar lugar a usos prácticos como el desarrollo de aplicaciones en teléfonos móviles para realizar esta detección, o la creación de dispositivos diseñados para prevenir la violencia doméstica a través de la voz y la biometría.

Apéndice A

Apéndice A: Desglose de resultados

A.1. Resultados de Precision y Recall

A continuación se recogerán los resultados complementarios de las pruebas realizadas en el Capítulo 5, el que solo se mostraban los valores correspondientes a F Score. Para cada una de ellas, se presentarán en primer lugar los valores obtenidos sobre Precision y, posteriormente, los correspondientes de Recall.

Pruebas dependientes del locutor

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.921	0.921	0.925	0.883
SVM linear	0.895	0.899	0.925	0.917
SVM poly	0.920	0.930	0.964	0.921
SVM rbf	0.942	0.949	0.972	0.972
SVM sigmoid	0.813	0.815	0.865	0.853

CUADRO A.1: Precision prueba dependiente del locutor Set 1

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.884	0.891	0.939	0.934
SVM linear	0.740	0.721	0.852	0.832
SVM poly	0.833	0.851	0.922	0.882
SVM rbf	0.876	0.873	0.913	0.882
SVM sigmoid	0.617	0.611	0.722	0.681

CUADRO A.2: Recall prueba dependiente del locutor Set 1

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.891	0.896	0.933	0.939
SVM linear	0.849	0.860	0.880	0.903
SVM poly	0.920	0.922	0.958	0.954
SVM rbf	0.927	0.930	0.957	0.958
SVM sigmoid	0.769	0.775	0.717	0.763

CUADRO A.3: Precision prueba dependiente del locutor Set 2

	Nivel 1 umbral 1	**Nivel 1 umbral 2**	**Nivel 2 umbral 1**	**Nivel 2 umbral 2**
MLP	0.831	0.843	0.840	0.863
SVM linear	0.663	0.696	0.715	0.733
SVM poly	0.801	0.802	0.785	0.800
SVM rbf	0.794	0.800	0.780	0.813
SVM sigmoid	0.528	0.534	0.509	0.521

CUADRO A.4: Recall prueba dependiente del locutor Set 2

	Nivel 1 umbral 1	**Nivel 1 umbral 2**	**Nivel 2 umbral 1**	**Nivel 2 umbral 2**
MLP	0.891	0.897	0.935	0.939
SVM linear	0.817	0.837	0.876	0.891
SVM poly	0.906	0.914	0.935	0.935
SVM rbf	0.919	0.920	0.959	0.948
SVM sigmoid	0.731	0.749	0.730	0.774

CUADRO A.5: Precision prueba dependiente del locutor Set 3

	Nivel 1 umbral 1	**Nivel 1 umbral 2**	**Nivel 2 umbral 1**	**Nivel 2 umbral 2**
MLP	0.834	0.834	0.880	0.884
SVM linear	0.659	0.669	0.715	0.779
SVM poly	0.811	0.816	0.861	0.869
SVM rbf	0.808	0.813	0.836	0.856
SVM sigmoid	0.497	0.488	0.430	0.550

CUADRO A.6: Recall prueba dependiente del locutor Set 3

Prueba independiente del locutor

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.737	0.759	0.748	0.773
SVM linear	0.710	0.736	0.744	0.783
SVM poly	0.736	0.760	0.744	0.768
SVM rbf	0.738	0.764	0.746	0.767
SVM sigmoid	0.708	0.728	0.705	0.754

CUADRO A.7: Precision prueba independiente del locutor entre Set 1 y Set 2

	Nivel 1 umbral 1	Nivel 1 umbral 2	Nivel 2 umbral 1	Nivel 2 umbral 2
MLP	0.925	0.938	0.934	0.946
SVM linear	0.627	0.627	0.741	0.757
SVM poly	0.817	0.830	0.911	0.930
SVM rbf	0.898	0.916	0.915	0.919
SVM sigmoid	0.632	0.618	0.601	0.666

CUADRO A.8: Recall prueba independiente del locutor entre Set 1 y Set 2

A.2. Correspondencia eje gráfica - ID

En esta sección se recogen en tablas las correspondencias entre los valores del eje X de las gráficas del Capítulo 5 Sección 5.3

Número eje x	Hombres	Mujeres
1	652033332	62963719
2	1015666824	935941053
3	1397020749	1395228143
4	1420900415	2054751935
5	1739028311	92305089
6	513604950	304102792
7	852630991	334844205
8	902398068	
9	1143102813	
10	1458206716	
11	1626125349	
12	1686645257	
13	1756953694	
14	1777108864	

CUADRO A.9: Correspondencia género gráfica - ID

Número eje x	**Set 1**	**Set 2**	**Set 3**
1	62963719	12782919	62963719
2	652033332	49425811	652033332
3	935941053	92305089	935941053
4	1015666824	304102792	1015666824
5	1395228143	334844205	1395228143
6	1397020749	513604950	1397020749
7	1420900415	852630991	1420900415
8	1739028311	902398068	1739028311
9	1777769661	1143102813	1777769661
10	2054751935	1458206716	2054751935
11		1626125349	12782919
12		1686645257	49425811
13		1756953694	92305089
14		1777108864	304102792
15			334844205
16			513604950
17			852630991
18			902398068
19			1143102813
20			1458206716
21			1626125349
22			1686645257
23			1756953694
24			1777108864

CUADRO A.10: Correspondencia sets gráfica - ID

Apéndice B

Apéndice B: Código proyecto

Con el fin de compartir el proyecto con la comunidad científica, se ha decidido publicar el código utilizado durante el proyecto en el siguiente repositorio GitHub:

`https://github.com/minguezalba/Stress_Detection`

A continuación se muestra el archivo README adjunto al proyecto.

Stress Detection in Voice Signals

This project consists on detecting stress in voice signals based on heart rate values. The project is divided in three main diferent parts: data processing, feature extraction and classification. We are going to use VOCE database so the code will be focused on its structure. However, individual functions can be used separately. We strongly recommend to read the paper about this project before trying to run it.

Getting Started

You will need some requisites to run this project.

Prerequisites

- Download database VOCE from its own webpage
- MATLAB 2016
- MIRtoolbox
- VOICEBOX: Speech Processing Toolbox for MATLAB
- Python 2.7
- Spyder IDE from Anaconda Python
- Sklearn version: scikit-learn 0.18.1

Code Structure and Usage

1. Data Processing

From raw files in VOCE database you will need to follow some steps in order to have usefull data. You will find the scripts in "Procesado BBDD"folder.

- Use script_ rename.bat file to rename negative IDs so all of them are positive numbers
- Use analisis_ archivos.m to analysis how much data is available (wav files, sensor files...) for each ID
- Extract Heart Rate values from sensors xml files with analisis.xml function. Check both Zts and Zecg values are similar with ejemplos_ Zecg_ Zts.m script.
- Select the files you will use based on mean and standar deviation from ejemplos_ pre_ baseline.m script.
- Extract the heart rate values you will use to generate labels with crear_ Zecg.m script.

2. Feature Extraction with MATLAB

In this stage we will need Heart Rate values from sensors files and their corresponding speech wav files. We will extract some basic features and their statistics with some tools like MIRtoolbox and VOICEBOX for MATLAB. We will also create the corresponding labels (1 = stress, 0 = no stress) for each segment from the audio file based on the Heart Rate values.

You just need to run main.m script in Matlab and you will get feature matrix and labels for each ID.

3. Classification with Python

When we already have feature matrix and labels for each file, we well go on with the classification stage.

- First of all, we well create train and test sets with conjuntos_ depend.py or conjuntos_ independ.py depending on the kind of test we want to do. We can try speaker dependent or speaker independent test.
- Once you have X_ train, X_ test, Y_ train and Y_ test sets, you will be able to run the classification with classifier.py script. You can try different classifiers with different parameters.
- Classifiers are evaluated with precision, recall and f scores.
- You can get some comparative graphics about F score and Heart Rate values for each ID with graficas.py script.

Author

Alba Minguez Sanchez, June 2017

Apéndice C

Apéndice C: English Summary

Chapter 1: Introduction

Introduction

Emotions are a fundamental part of the human being because of their great influence on perception and everyday tasks, such as learning, communication and even rational decision making.

Thanks to the technological advances of the last years, affective computation is born. This science, among other things, is responsible for analyzing the emotions in the speech signals. Therefore, we can measure empirically how changes in a person's autonomic nervous system indirectly affect a person's speech.

Stress detection can be used in numerous applications: jobs as pilots, air traffic controllers, emergency personnel; Public events such as job interviews and speeches; or even in cases to detect domestic violence.

In spite of this multiple situations, recent research studies have been done under laboratory conditions. The results obtained in these situations provide a limited generalization to evaluate real stress situations. So it is desirable to record those observations in real stress situations.

In our study, we addressed stress detection for a group of students who gave public speeches during which physiological data was taken. The main goal of this project is to teach a machine to perform stress detection solely on speech signals. To do this, we base the stress annotation on heart rate values and extract a basic feature set from speech in order to achieve discrimination between stress and non-stress situations.

Motivation and Goals

As mentioned above, the main objective of this study is to design and test a stress detection system from speech signals. The motivation behind this goal is based on the need to design a detector accurate enough for real stress situations.

However, the project focuses on the fulfillment of particular objectives that help to achieve the general purpose.

First of all, one of the intentions of this study is to find how strong the relationship between speech and stress is from the speaker and his heart rate. It is also of interest to look for differences based on gender and age during periods of stress, and to help answering the question of how audio quality affects the detection.

Our study will also test different machines for detection in order to find which systems provide better results. Various data sets will also be tested both in training and testing stages.

Chapter 2: State of the Art

Stress definition

According to the definition proposed by Murray et al. "Stress is the observable variability in certain speech features due to a response to stressors..On the other hand, Hansel points to the difficulty of separating it from the rest of the emotions, since it usually appears as a combination of them, like anger or sadness.

In our study we will analyze acute stress, that is, the one that appears during a short term in demanding situations. This is usually followed by an overexcitation of the nervous system (sweating, blood pressure increase...)

Speech technologies

Speech technologies are the set of techniques, methods and algorithms designed to model human - machine communication from oral language (human voice). Down below most commonly used technologies are briefly defined:

- **Speech Recognition (SR)**. It consists on the science that is responsible for processing speech signal of the human being and recognizing the information contained in it, essentially the text. Among other things, speech recognition systems can be classified as speaker dependent or independent.
- **Speech Synthesis.** This is the artificial production of human voice with the aim of making it sounds as natural as possible.
- **Speaker Recognition** This technique consists on identifying the person speaking from the biometric features from his speech.

Stress empirical description

Environments

Nowadays, studies about emotions or stress recognition have followed several ways to obtain voice recordings in different stress situations. Depending on where the recording takes place, it can be performed in a real event or in a laboratory. Within this latter category, stress can be interpreted or induced.

Databases

There are many databases that collect audio data and sensors on emotions and stress. One of the best known is SUSAS, a database that contains data about aircraft communications in different stress situations.

However, in our study we will use the so called VOCE database which collects audios from speeches by 45 students from the University of Porto and their corresponding data on heart rate at different times of recording. In addition to this, you also have personal information about each participant. We will use this dataset because it provides us data taken in real events and, also, it contains different recording times which will let us calculate the decision thresholds.

Annotation and biometrics

Different biometric and psychological measures have been used to detect stress in speech.

Measures related to heart rate, such as R peaks or RR intervals are some of the most common. Galvanic Skin Response is also used, based on the changes of electricity and temperature of the skin. Blood pressure and electromyography, which analyzes the electrical activity of muscles, are other possible measures.

Psychological tests are also used to measure certain aspects of emotional stress, such as the State-Trait Anxiety Inventory (STAI).

Feature Extraction Methods

In order to understand thoroughly what a speech system is, a block diagram with the elements that compose it is represented in the Figura 2.7.

Feature extraction algorithms are those algorithms and techniques whose goal is to calculate a set of feature vectors which provide a compact representation of the most important aspects of the input data to the system, in our case, the speech signals.

One the one hand, features such as pitch and frequencies of the formants can be calculated with LPC technique. The main idea on which LPC is based is that a speech sample can be approximated as a linear combination of previous speech samples.

On the other hand, spectral analysis is another of the most common techniques. Thanks to this, we obtain the MFCC, values that are used to model the human auditory perception.

There are some other features such as jitter and shimmer, which measure variability in amplitude and time, as well as speech rate, that is, the speed at which the speech is uttered.

Classifiers

In our project, we will focus on supervised learning algorithms, that is, the ones which use a set of already labeled samples (training set) to train and, then, try to assign a label to a set of unknown samples (test set).

Multilayer Perception (MLP) is perhaps the most popular algorithm for both classification and regression. MLP is an Artificial Neural Network (ANN) that is often composed of several layers of nodes with unidirectional connections, often trained by backpropagation.

Another of the classifiers most used in these technologies are the SVM (Support Vector Machine). The SVM classifiers are mainly based on the use of nonlinear functions that map the original features in a larger space dimension, thus allowing classification using a linear classifier.

Chapter 3: Design

Problem definition

The problem that we are trying to solve is the detection of stress in speech signals, that is, to be able to determine if a sample of speech signal presents stress or not. To do this, we will use VOCE dataset. So a set of voice recordings and their heart rate values will be available.

The data is clustered in subsets according to the quality of the audios. Subsequently, their features are extracted at different analysis levels. Labels or targets are generated from their physiological data based on different decision thresholds which will be determinated.

Finally, the goal is to design several detection machines to test all the data subsets at the different analysis levels to draw the corresponding conclusions.

Solution design

First of all, feature extraction from the audio recordings, as well as generation of the labels from the HR data are performed. Both processes are done at two levels or segment sizes, one smaller and one larger (segmental and suprasegmental), which will be calculated from the smaller. In addition to this, labels are designed based on two decision thresholds calculated from heart rate values.

For feature extraction, both LPC and MFCC extraction techniques are used, from which we obtain a set of features and their corresponding statistics about pitch, formants frequencies, MFCC coefficients and signal energy.

After obtaining this data, we move on to the design phase of the classification machine. To do this, we divide, in the speaker dependent tests, our data into training sets (80 %) and test sets (20 %). Machines are designed with supervised MLP and SVM learning algorithms using different Kernel functions for the latter. We generate both speaker dependent and independent models and evaluate our classifier for the different cases.

Chapter 4: Implementation

Processing the database and creating sets

Firstly, all available files in the database were listed. In this way, it was observed that out of the total of 45 IDs or participants that comprise it, 33 contained the three audio files corresponding to the three different recording instants *(prebaseline, baseline and recording)* and their HR values, while 12 of them were incomplete.

After this, data extraction from sensors files containing values related to heart rate started. This extraction was done using MATLAB. These files consist of two different values: *Zecg* and *Zts*. Zecg are mean values of HR, while Zts are the time instants in which R peaks occur on the electrocardiogram.

Each of these values was analyzed and transformed separately into useful values for our project. On the one hand, formatting errors of Zecg values were corrected. On the other hand, Zts values were processed until they were in bpm units so we could compare them with Zecg.

After comparing Zecg values with Zts, it was observed that Zecg corresponded to mean and filtered values of Zts. Taking into account this, and observing that not all of the IDs had Zts values, we decided to use only Zecg values in the project, which would later be used to calculate the decision thresholds of the labels.

Respect to the creation of sets, some of the IDs were firstly discarded because they did not meet some minimum quality criteria. Subsequently, 2 sets of data were created: Set 1 with the best audios, and Set 2 with the remaining ones. In test chapter we also used Set 3 which is the union of both.

Processing speech signals

The processing of speech signals consists of the following sequence of steps:

Firstly, we decided to pass audio recordings from stereo to mono. Subsequently, the sampling frequency is reduced, from 44100 Hz to 16000 Hz.

The next step is to compare signals with others, that is, normalization. In this phase, we required that all signals had their amplitude in a range of [-1,1] with mean 0.

The last step is possibly the most important of the pre-processing of the signals. In this phase a Voice Activity Detector (VAD) was designed with the objective of detecting the signal silence intervals and eliminating them, since they do not contain useful information for our project. This segmentation is done in blocks of 1 second so as not to lose synchrony between audios and Zecg values used for the labels.

Generating labels

Before generating the labels, we will set the two levels of analysis explained previously.

Our proposal is to work on the following levels of analysis:

- **Level 1:** a 2 second window size with 1 second shifting.
- **Level 2:** five times the sizes of level 1, that is, window size of 10 seconds with 5 seconds shift.

Respect to the generation of labels, the first step is to establish the decision thresholds, that is, choose how we determine when, in an audio sample, there is stress or not.

In our study we propose two different thresholds, unique for each ID, both based on the values of HR from *base* instant. The first threshold, threshold 1, will be based on the value resulting from adding the standard deviation to the mean of the HR values. The second, called threshold 2, establishes a percentage of frames that will be considered as stressed and therefore correspondes to the 75 % percentile, in this particular case.

Therefore, we apply the criterion that if the HR value of a sample exceeds the threshold, the corresponding label will be 1 (stress) and, conversely, if the value is below the threshold, the label will be 0 (no stress). In this way, we will obtain the labels for level 1 that will be used to later calculate those of level 2. Before calculating the latter, the samples corresponding to silences in the audios will be eliminated.

Feature Extraction

Our project proposes the extraction of a basic set of features as well as of its statistics in order to obtain enough information for the detection. This stage has been done with MATLAB and the VoiceBox toolbox.

Pitch, frequencies of the formants and MFCCs have been extracted with an analysis window of 20 ms. Subsequently, the mean and variance statistics for window level 1 were calculated. In addition to these features, we estimate the energy of the speech signal.

After calculating all feature vectors and stacking them one after another, we obtain an array of dimensions 33xN1, where N1 is the number of frames corresponding to the first window level. Finally, we calculate the feature matrix at level 2. To do this, we will calculate the characteristic vectors by averaging the values of level 1.

Classification and evaluation

In this last stage we use all the information previously generated as input to our classifier. For each machine, 4 possible combinations between features and labels will be evaluated, 2 in each level. This part of the project has been developed with the Python programming language.

In the case of the speaker dependent tests, the data will be randomly divided into two sets: train (80 %) and test (20 %). After this, the training set is

balanced by cloning of samples of the minority class. Subsequently, train and test sets are normalized.

For the design of the machines we decided to use the default values from Python *sklearn* package for both MLP and SVM. Once we trained the machines with the training data, we predict the labels for the test set.

To evaluate the performance of our machine Precision, Recall and F Score metrics are computed. The criterion with which we will evaluate our classifier will be, mainly, F Score value, so that the closer it is to 1, the better our ranking will be.

Chapter 5: Planning and Budget

Regarding the temporary execution of the project, it has been developed in an approximate period of 20 weeks, exactly from February 10, 2017 to June 15, 2017.

As for the budget, after analyzing the physical, software and human resources, a total project budget of six thousand three hundred and eighty-six euros and eighty-one cents without VAT.

Chapter 6: Testing, results and conclusions

In order to design an accurate enough detector for real stress events, our study has made multiple combinations between feature extraction phase and classification stage, in order to obtain, from all of these combinations, the best classification score.

Firstly, two large groups of tests were done: speaker dependent and independent tests. The results of these tests demonstrate several aspects about the different levels and thresholds used. As for the levels, it has been shown that a larger window results in higher detection scores, as expected, regardless of the set and the speaker dependence or independence partition. However, regarding the detection thresholds this uniformity is not observed. Although the difference is not very large, threshold 1 gives better results for the set of higher quality data, ie set1, while threshold 2 works best in those sets whose quality is lower, such as set 2 And set 3.

Regardless of the test, audio quality has been shown to affect the detection of stress. The best results were obtained for Set 1, whose audios were of higher quality, while the lowest values were given for Set 2, of worse quality; and intermediate values for Set 3, which combines samples from both.

As for machine design, it has been found that, in general, SVM classifiers with Kernel RBF function give better results for small sample sets, around 85-94 %, unlike the MLP classifier, whose performance increases for larger sets, around 86-93 %. However, the difference between these two classifiers is small, so we should calculate both confidence margins to determine if there is any significant difference between them.

The comparison between speaker dependent and independent tests answers the question of the degree of dependence in the classification stage. Considering that the results are slightly higher in the dependent tests, we can affirm that there is some dependence on the speaker in our system. However, the results obtained are not too different for this to be a determining factor in our project.

Another interesting goal of our study was to find a link between heart rate values and F Score. For this purpose, comparative graphs were generated from which several conclusions have been drawn.

Firstly, when we perform the detection at level 1, we find a relationship between the heart rate difference between the recording and base samples from each ID with their individual F score. Apparently, those IDs whose difference is lower get lower F scores too. However, at level 2 this dependency disappears.

The goal of these graphs was, as well, to look for differences between gender and ages of the different participants. However, no significant evidence has been observed in either case, possibly because of the limited information and quantity of samples in the selected database.

In summary, it is concluded that, in general terms and after analyzing the achievement of the objectives, the project has been successfully completed after quite accurately achieving the detection of stress in speech signals in real events.

Chapter 7: Future Work

Once the study is completed, new lines of work emerge to continue, analyze or fix individual aspects of the project in depth.

One of the classic ways would be to extend the variety of features to be extracted from audios, such as: jitter, shimmer, speech rate, etc. The extraction of new features would allow to study those that are more correlated with stress detection.

In order to improve the quality of speech signals for the feature extraction phase, another point to correct would be to preprocess those audios that contain saturation due to the microphone. This would significantly influence the stage of normalization of audios, eliminating confusions between a maximum of saturation and a real one.

In this same phase of preprocessing of audios, another aspect to enhance would be the elimination of silences in audios. In our case, this elimination has been done separately both in audio files and labels, trying not to lose synchronism, but performing a block-based segmentation, sometimes inaccurate, which may add include errors. To correct this, one of the possible solutions would be to add a third level of analysis that encompassed both audio and labels vectors and to jointly eliminates the silences of both with more precision.

In order to improve the classification, a further adjustment of the parameters of the classifiers should be done, since we have used the default values.

To do this, it would be necessary to retain a validation set or perform cross-validation for each parameter to be adjusted. On the other hand, in view of the small differences that exists between the MLP and SVM rbf classifier results, it would be interesting to calculate their confidence intervals to determine whether they are statistically significant.

Finally, and for longer term projects, stress detection based on speech features and biometric measures could lead to practical applications such as their embedding in mobile phones to perform this detection, or the design of devices to prevent domestic violence through voice and biometrics.

Bibliografía

[1] R. Picard, *Affective Computing*. MIT Press, 2000, ISBN: 9780262661157. dirección: https://books.google.es/books?id=GaVncRTcb1gC.

[2] Y. J. Zanstra y D. W. Johnston, «Cardiovascular reactivity in real life settings: Measurement, mechanisms and meaning», *Biological Psychology*, vol. 86, n.º 2, págs. 98 -105, 2011, Cardiovascular Reactivity at a Crossroads: Where are we now?, ISSN: 0301-0511. dirección: http://www.sciencedirect.com/science/article/pii/S0301051110001353.

[3] (2017). Ethics, The EU Framework Programme for Research and Innovation, dirección: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/ethics (visitado 04-06-2017).

[4] (2016). EUROPEAN CIVIL LAW RULES IN ROBOTICS, dirección: http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf (visitado 29-05-2017).

[5] (2017). VOCE Corpus database, dirección: http://cloud.futurecities.up.pt/~voce/metadata/ (visitado 27-05-2017).

[6] A. Aguiar, M. Kaiseler, H. Meinedo, T. E. Abrudan y P. R. Almeida, «Speech Stress Assessment using Physiological and Psychological Measures», 2013. dirección: http://paginas.fe.up.pt/~voce/docs/mcss05-aguiar.pdf.

[7] I. R. Murray, C. Baber y A. South, «Towards a definition and working model of stress and its effects on speech», *Speech Communication*, vol. 20, n.º 1, págs. 3 -12, 1996, ISSN: 0167-6393. dirección: https://goo.gl/CVXU5X.

[8] S. P. John H. L. Hansen, «Speaker Classification I», en. Springer Berlin Heidelberg, 2007, cap. 6, 108–137. dirección: https://link.springer.com/chapter/10.1007%2F978-3-540-74200-5_6.

[9] L. Miller, A. Smith y L. Rothstein, *The Stress Solution: An Action Plan to Manage the Stress in Your Life*. Pocket Books, 1994, ISBN: 9780671753115. dirección: https://books.google.es/books?id=3y1OcoaBl9AC.

[10] B. Gold, N. Morgan y D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 2.ª ed. Wiley-Interscience, 2011, ISBN: 978-0-470-19536-9.

[11] D. Jurafsky, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2009, ISBN: 9780135041963.

[12] (2016). Speech technology, dirección: https://en.wikipedia.org/wiki/Speech_technology (visitado 04-06-2017).

[13] D. Ververidis y C. Kotropoulos, «Emotional speech recognition: Resources, features, and methods», *Speech Communication*, vol. 48, n.º 9, págs. 1162 -1181, 2006. dirección: http://www.sciencedirect.com/science/article/pii/S0167639306000422.

[14] T. Vogt, E. André y J. Wagner, «Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation», *Affect and Emotion in HCI, LNCS 4868*, 75–91, 2008. dirección: https://goo.gl/NgJzc0.

[15] (1999). SUSAS Database, dirección: https://catalog.ldc.upenn.edu/LDC99S78 (visitado 26-05-2017).

[16] A. Aguiar, M. Kaiseler, M. Cunha, J. Silva, H. Meinedo y P. R. Almeida, «VOCE Corpus: Ecologically Collected Speech Annotated with Physiological and Psychological Stress Assessments.», *Language Resources and Evaluation Conference*, 2014. dirección: http://paginas.fe.up.pt/~voce/docs/lrec2014-aguiar.pdf.

[17] R. Fernández y R. W. Picard, «Modeling Drivers' Speech Under Stress», *Speech Communication*, págs. 145-149, 2003. dirección: http://affect.media.mit.edu/pdfs/03.fernandez-picard.pdf.

[18] D. Ververidis y C. Kotropoulos, «A Review of Emotional Speech Databases», dirección: http://delab.csd.auth.gr/bci1/Panhellenic/560ververidis.pdf.

[19] G. Demenko, «Voice stress extraction», *Proceedings of the Speech Prosody 2008 Conference*, págs. 53-56, 2008. dirección: https://goo.gl/o2sVd1.

[20] G. Demenko y M. Jastrzębska, «Analysis of Voice Stress in Call Centers Conversations», *Speech Prosody, 6th International*, 2012. dirección: http://isle.illinois.edu/sprosig/sp2012/uploadfiles/file/sp2012_submission_234.pdf.

[21] M. D. Julião, «Feature Sets for Stressed Speech Discrimination», 2014. dirección: http://paginas.fe.up.pt/~voce/docs/teseMarianaJuliao.pdf.

[22] (2017). Imagen picos R, dirección: https://goo.gl/hejNyf (visitado 12-06-2017).

[23] G. Nishigawa, N. Natsuaki, Y. Maruo, M. Okamoto y S. Minagi, «Galvanic skin response of oral cancer patients during speech», *Journal of Oral Rehabilitation*, vol. 30, n.º 5, págs. 522-525, 2003, ISSN: 1365-2842. dirección: http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2842.2003.01131.x/full.

[24] H. Kurniawan, A. V. Maslov y M. Pechenizkiy, «Stress detection from speech and Galvanic Skin Response signals», en *CBMS*, 2013. dirección: https://goo.gl/N5qOfc.

[25] K. S. Lee, «EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables», *IEEE Transactions on Biomedical Engineering*, vol. 55, n.º 3, págs. 930-940, 2008, ISSN: 0018-9294. dirección: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4454043.

[26] (2013). Inventario de ansiedad estado-rasgo STAI, dirección: http://www.psico-system.com/2013/05/inventario-de-ansiedad-estado-rasgo-stai.html (visitado 27-05-2017).

[27] (). Imagen formantes, dirección: https://goo.gl/4UcN0U (visitado 12-06-2017).

[28] (2017). Imagen MFCC, dirección: https://goo.gl/NwnCFy (visitado 12-06-2017).

[29] N. Sundaram, B. Y. Smolenski y R. E. Yantorno, «INSTANTANEOUS NONLINEAR TEAGER ENERGY OPERATOR FOR ROBUST VOICED – UNVOICED SPEECH CLASSIFICATION», 2003. dirección: https://goo.gl/XV5b4I.

[30] J. P. Teixeira, C. Oliveira y C. Lopes, «Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters», *Procedia Technology*, vol. 9, págs. 1112 -1122, 2013, ISSN: 2212-0173. dirección: http://www.sciencedirect.com/science/article/pii/S2212017313002788.

[31] H. Palo y M. N. Mohanty, «Reduced Feature Extraction for Emotional Speech Recognition», 2015. dirección: https://www.researchgate.net/publication/303280369_Reduced_Feature_Extraction_for_Emotional_Speech_Recognition.

[32] E. Zarrouk, Y. Ben Ayed y F. Gargouri, «Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study», *International Journal of Speech Technology*, vol. 17, n.º 3, págs. 223-233, 2014, ISSN: 1572-8110. dirección: http://dx.doi.org/10.1007/s10772-013-9221-5.

[33] R. Schwartz, J. Klovstad, J. Makhoul y J. Sorensen, «A preliminary design of a phonetic vocoder based on a diphone model», en *ICASSP '80. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 1980, págs. 32-35. DOI: 10.1109/ICASSP.1980.1171037.

[34] (2017). Imagen diagrama HMM, dirección: https://www.researchgate.net/figure/278411252_fig3_Figure-27-A-4-state-Left-Right-HMM-model (visitado 12-06-2017).

[35] E. Zanaty, «Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification», *Egyptian Informatics Journal*, vol. 13, n.º 3, págs. 177 -183, 2012, ISSN: 1110-8665. dirección: http://www.sciencedirect.com/science/article/pii/S1110866512000345.

[36] (2017). Imagen diagrama MLP, dirección: https://goo.gl/xO7ZmC (visitado 12-06-2017).

[37] (2017). Imagen sistema de habla, dirección: http://www.cc.gatech.edu/~athomaz/classes/CS8803-HRI-Spr08/icat/files/base.png (visitado 12-06-2017).

[38] O. Simantiraki, G. Giannakakis, A. Pampouchidou y M. Tsiknakis, «Stress Detection from Speech Using Spectral Slope Measurements», *Conference: 6th EAI International Symposium on Pervasive Computing Paradigms for Mental Health At Barcelona*, 2016. dirección: https://goo.gl/9ELTEV.

[39] I. MOHINO, R. GIL-PITA y L. A. PÉREZ, «Stress detection through emotional speech analysis», 2012. dirección: http://www.wseas.us/e-library/conferences/2012/Prague/ECC/ECC-35.pdf.

[40] R. Ma, «Parametric Speech Emotion Recognition Using Neural Network», 2014. dirección: http://www.diva-portal.org/smash/get/diva2:756207/FULLTEXT01.pdf.

[41] S. Hewlett, «Emotion Detection from Speech», 2007. dirección: http://cs229.stanford.edu/proj2007/ShahHewlett%20-%20Emotion%20Detection%20from%20Speech.pdf.

[42] B. V. Sathe-Pathak y A. R. Panat, «Extraction of Pitch and Formants and its Analysis to Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person», *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4*, 2012. dirección: https://www.ijcsi.org/papers/IJCSI-9-4-1-296-299.pdf.

[43] S. Gonzalez y M. Brookes, «PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, n.º 2, págs. 518-530, 2014.