Universidad Carlos III de Madrid

Escuela Politécnica Superior

Grado en Ingeniería de Sistemas Audiovisuales

Trabajo Fin de Grado

Extracción de características espectrales y prosódicas para reconocimiento de emociones

Autora: Irene M. Navidad Peñalba

Tutora: Ascensión Gallardo Antolín

Leganés, Julio de 2014

Título: **Extracción de características espectrales y prosódicas para reconocimiento de emociones**

Autora: **Irene M. Navidad Peñalba**

Tutora: **Ascensión Gallardo Antolín**

EL TRIBUNAL

Presidente: Bernardo D'Auria

Vocal: Sara Pino Povedano

Secretario: Eduardo Martínez Enríquez

Realizado el acto de defensa y lectura del Proyecto Fin de Grado el día 4 de Julio de 2014 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO PRESIDENTE

“Let’s not forget that the little emotions
are the great captains of our lives and
we obey them without realizing it.”

Vincent Van Gogh

Agradecimientos

Llegado este momento, me gustaría agradecer todo el apoyo prestado a todas las personas que han formado parte de mi vida durante todos estos años.

En primer lugar, a mis padres, Juan y Milagros, por inculcarme todos los valores a los que hoy día me aferro, por aguantar todos mis agobios, por saber darme consejos, por hacerme la vida un poco más fácil y porque sin ellos llegar hasta aquí no hubiera sido posible.

Quiero agradecerle encarecidamente todo su apoyo a Ascensión, mi tutora, por haberme dedicado tanto tiempo, por estar siempre presente cuando la he necesitado y porque sin su ayuda no habría conseguido realizar este proyecto.

A toda la gente que he conocido durante este largo camino y que me llevo para toda la vida. A Luis, por apoyarme incondicionalmente y estar a mi lado cada día. A Elena, por todos los años que llevamos juntas y porque nunca falla. A Curro, por hacerme más amenos los días eternos en la universidad. A Jesús, mi compañero de biblioteca.

A mis compañeros de clase: Marta, Raquel, Aroa, Guillermo, Víctor, Xandre, Carlos, Patricia, Andrea y Valle. A María, por amenizarme los últimos meses con sus divertidas anécdotas. A todas esas personas especiales que compartieron el erasmus conmigo.

A las mejores personas que una puede tener como amigas: Elena, Carmen y Esther. Gracias por aguantarme y por darme toda la energía positiva durante todos estos años, aunque los últimos sea a distancia.

A mis compañeras y sobre todo amigas del equipo: Jimena, Sara, Ángela, Rebeca y Cristina. A mis entrenadores Toni y Goyo, que me enseñaron que nada se consigue sin luchar y sufrir por ello.

A todos vosotros, y a todos los que olvido nombrar, GRACIAS de corazón.

Resumen

En las últimas décadas, los sistemas automáticos de reconocimiento de patrones han ganado mucha importancia debido al interés de crear interacciones entre el hombre y la máquina lo más naturales posibles.

Este trabajo fin de grado se centra en los sistemas automáticos de reconocimiento de emociones que, a partir de la voz de un hablante y usando técnicas de aprendizaje máquina, son capaces de reconocer el estado emocional del locutor. Este tipo de sistemas pueden ser muy útiles para mejorar la calidad de vida de las personas, especialmente para las que tienen algún tipo de discapacidad o incluso para mejorar investigaciones que están relacionadas con la emoción, como puede ser en el campo de la psicología o neurología.

El objetivo de este trabajo es diseñar e implementar en Matlab un sistema de reconocimiento automático de emociones. Para ello, previamente se han estudiado las bases teóricas y así poder comprender cómo se relacionan las emociones con los aspectos físicos y acústicos de la voz. Se han investigado algunas técnicas de clasificación para decidir cuál se adapta mejor a los objetivos de este proyecto.

El sistema desarrollado consta de dos etapas: parametrizador y clasificador. El primer módulo se encarga de la extracción de tanto características espectrales (coeficientes mel-cepstrales) como características prosódicas de la señal de voz (frecuencia fundamental, frecuencia del primer formante, parámetros de calidad acústica, duración). En la segunda etapa, se procede a la comparación de dichas características con los patrones de emociones obtenidos mediante un proceso de entrenamiento, a partir de una base de datos previamente etiquetada. Dicha comparación se realiza utilizando técnicas de clasificación basadas en máquinas de vector soporte.

Para determinar las prestaciones del sistema, se han realizado una serie de experimentos considerando distintas características espectrales, prosódicas y su combinación. A la vista de los resultados, se ha podido concluir que las características espectrales extraídas a nivel de clase contienen una información más precisa sobre las emociones que las extraídas a nivel de expresión, y que la combinación de dichas características espectrales con las prosódicas producen los mejores resultados de reconocimiento.

Palabras clave: Reconocimiento, emociones, voz, MFCC, prosodia, SVM.

Abstract

In recent decades, the automatic pattern recognition systems are gaining a lot of importance due to the interest of creating interactions between man and machine as natural as possible.

This final project is focused on the automatic emotion recognition, which from the voice of a speaker and using machine learning techniques, is able to recognize the emotional state of the speaker. These types of systems can be very helpful to improve the quality of life of people, especially those with some kind of incapacity or even to enhance researches that are related with emotion, such as in the field of psychology or neurology.

The object of this project is to design and implement automatic emotion recognition in Matlab. To do so, theoretic basis have previously been studied in order to understand how emotions relate to the physical and acoustic aspects of voice. Besides, some classification techniques have been treated in order to decide which one is the best to achieve the objectives of this project.

The developed system consists of two stages: parameter assignment and classifier. The first module is responsible for the extraction of both spectral characteristics (mel-cepstral coefficients) as prosodic characteristics of the voice signal (fundamental frequency, of the first formant frequency, sound quality settings, duration). In the second stage, we compare these characteristics with emotion patterns obtained through a training process from a data base pre-labeled. This comparison is performed using classification techniques based on support vector machines.

To determine the performance of the system, a number of experiments have been made considering different spectral and prosodic features and their combination. Regarding the results, it has been concluded that the spectral features extracted at class level contain more precise information about the emotions than the ones extracted at expression level, and that the combination of these spectral features with the prosodic ones produce the best recognition results.

Keywords: Recognition, emotions, voice, MFCC, prosody, SVM.

Índice general

Índice de figuras

Índice de tablas

Índice de ecuaciones

Capítulo 1

Introducción y objetivos

La función de este capítulo introductorio es situar al lector en el marco tecnológico que ha dado lugar a la realización de este proyecto. Por otra parte, se explica el porqué de este trabajo y se describen los objetivos que se quieren alcanzar.

En los siguientes capítulos se describe de una forma más precisa la base teórica del sistema propuesto, las herramientas utilizadas para su implementación, así como la experimentación realizada y las principales conclusiones extraídas de este trabajo.

1.1. Marco Tecnológico

Cada día aparecen nuevas tecnologías cuyo objetivo es hacer más fácil la vida del ser humano. Estas tecnologías están en continuo desarrollo, lo cual se puede observar, por ejemplo, en la telefonía o en la informática, cuyas aplicaciones quedan obsoletas en cuestión de meses.

Actualmente, podemos encontrar en distintos aparatos tecnológicos algunas aplicaciones de reconocimiento de voz, como por ejemplo la escritura de mensajes y textos a través del habla, o la llamada automática por voz, que consiste en decir el

nombre de la persona que se quiera llamar para que el dispositivo marque el teléfono que corresponde.

Uno de los últimos objetivos que se persigue en la actualidad consiste en conseguir una comunicación natural y sin esfuerzo en la interacción del hombre con la máquina. Los científicos e ingenieros que tratan de construir aparatos y sistemas avanzados se enfrentan a uno de los retos más complejos, imitar las habilidades naturales del ser humano. Resulta muy complicado desarrollar máquinas que caminen con naturalidad, que sean capaces de entender el habla, de expresar emociones, que puedan reconocer imágenes o que posean articulaciones semejantes a las del ser humano. Pero sin duda, la tarea más compleja es interpretar e imitar correctamente las capacidades del ser humano [1].

Las nuevas tecnologías permiten que surjan distintas formas de comunicación entre un ordenador y el usuario que lo maneja, más allá de la interacción que existe a través del ratón y el teclado. Para que el ordenador consiga establecer una interacción adecuada ha de ser capaz de tener alguna percepción del estado emocional del ser humano con el que interacciona, es decir, tiene que comprender sus emociones. Dos de los canales más informativos para la percepción de emociones por parte de una máquina son las expresiones faciales obtenidas a partir de un video y las expresiones léxico-fonéticas obtenidas de un discurso. Este trabajo se centra en identificar emociones a través de las expresiones léxico-fonéticas.

1.2. Motivación

La motivación principal de este trabajo es seguir avanzando en la investigación del reconocimiento de emociones en la voz humana para el futuro desarrollo de sistemas que pudieran ser empleados, no sólo para facilitar el uso diario de muchas aplicaciones, sino también para un público discapacitado, como puede ser el colectivo invidente o personas con problemas auditivos.

El reconocimiento de emociones a partir de la señal de voz es una disciplina que está ganando importancia en la interacción hombre-máquina. Tiene como objetivo identificar automáticamente el estado emocional o físico del ser humano a través de su voz. A los estados emocionales y físicos del locutor se les conoce como aspectos emocionales de la voz y forman parte de los aspectos paralingüísticos del habla. Aunque el estado emocional no altera el contenido lingüístico, éste es un factor importante en la comunicación humana, ya que proporciona más información que la meramente semántica acerca del interlocutor.

Se pueden distinguir dos canales en la comunicación humana. El primero, se ocupa de transmitir un mensaje de forma explícita, expresando el contenido semántico. El otro canal es el que permite una comunicación humano-humano. A través de este canal se transmite información implícita como el estado emocional del interlocutor, sexo, edad, etc.; es decir, se puede obtener información sobre la persona con la que se está realizando una comunicación. El reconocimiento automático de emociones se centra en este tipo de canal.

Con el progreso de las nuevas tecnologías y la introducción de sistemas interactivos, se ha incrementado enormemente la demanda de interfaces para comunicarse con las máquinas. Se puede ver un ejemplo de la interacción de una persona con una máquina en los teléfonos de atención al cliente de los principales operadores, donde la máquina puede generar respuestas más apropiadas si conoce el estado emocional de su interlocutor. La máquina recoge la respuesta del interlocutor y analiza su estado emocional, y dependiendo de los resultados obtenidos ésta debe ser capaz de cambiar de estrategia con el hablante para obtener unos resultados más satisfactorios tanto para el usuario, produciendo una sensación de satisfacción al ser comprendido más rápidamente, como para la máquina, obteniendo la información con menor dificultad y mayor rapidez.

1.3. Objetivos

El objetivo principal de este trabajo fin de grado es el diseño e implementación de un sistema de reconocimiento automático de emociones basado en la extracción de características espectrales y prosódicas de la señal de voz. En concreto, el sistema debe ser capaz de reconocer la emoción del hablante a partir de su voz, sin importar el contenido semántico de la oración expresada.

Para la realización de este trabajo se han tenido en cuenta los experimentos realizados en el estudio que presentaron Dimitri Bitouk, Ragini Verma y Ani Nenkova en el documento “Class-level features for emotion recognition” [2].

La finalidad de este trabajo fin de grado es, por tanto, adquirir los conocimientos necesarios para diseñar un sistema que cumpla las características presentadas en dicho estudio e implementarlo sobre el lenguaje de programación Matlab. Para ello será necesario el estudio de las características de las emociones, así como la investigación de técnicas y algoritmos que ayuden a su reconocimiento. De esta forma se desarrollará un sistema de clasificación de emociones y se evaluará su funcionamiento sobre una base de datos de voz expresiva estándar, con el objeto de determinar sus prestaciones y limitaciones.

1.4. Organización de la memoria

Para que la lectura de la memoria se pueda realizar de una forma sencilla, a continuación se incluye un breve resumen de cada capítulo.

En el capítulo 1 se hace una pequeña introducción al marco tecnológico del reconocimiento automático de emociones, seguido de la motivación que lleva a realizar este trabajo y de los objetivos que se quieren cumplir.

En el capítulo 2 se explican las teorías fundamentales que ayudan al entendimiento del sistema que se ha implementado y que se describe en capítulos posteriores. Para ello se habla de las características del habla, el sistema de producción vocal, la naturaleza de las emociones y se nombran diferentes bases de datos que han sido generadas para el desarrollo de este tipo de experimentos. Incluye también una descripción de algunos métodos de clasificación como pueden ser los modelos ocultos de Markov ("Hidden Markov Model", HMM) y las máquinas de vector soporte ("Support Vector Machines", SVMs).

En el capítulo 3 se describe el sistema automático de reconocimiento de emociones que se ha implementado, haciendo especial énfasis en la extracción de los diferentes tipos de características espectrales y prosódicas de la voz.

En el capítulo 4 de esta memoria se describen uno a uno los diferentes experimentos que se han desarrollado, incluyendo pruebas variadas que dan lugar a distintos resultados.

En el capítulo 5 se realiza un análisis económico de los costes que supondrían implementar un sistema de estas características. También se incluye una planificación de las tareas llevadas a cabo.

Para finalizar, se discute sobre los resultados obtenidos y se comparan para llegar a una conclusión sobre qué características acústicas son las más efectivas, indicando qué mejoras podrían implementarse.

Capítulo 2

Estado del arte

El objetivo de este capítulo es introducir al lector en el campo de investigación del reconocimiento automático de emociones y proporcionarle los conocimientos necesarios para poder entender el desarrollo y las conclusiones de este trabajo. Para ello, se explicarán las características más representativas de la señal de voz, cómo se produce, y las emociones que el ser humano es capaz de expresar a través de la voz.

Se hará un breve repaso sobre las diferentes bases de datos que se han desarrollado para este tipo de experimentos y se explicará en detalle la base de datos utilizada, *Berlín Data Base*. Además se repasarán brevemente los diferentes métodos de clasificación de datos que existen y que pueden aplicarse al reconocimiento de emociones.

Para finalizar el capítulo se exponen algunas aplicaciones actuales que se han desarrollado y que se usan en la vida cotidiana.

2.1. Introducción

El reconocimiento de emociones en la voz humana lleva siendo investigado desde hace varias décadas. A mediados de los años ochenta se empezaron a realizar estudios donde se utilizaban las propiedades estadísticas de algunas características acústicas. Más tarde, en los noventa, gracias a la evolución de los ordenadores se pudieron implementar algunos algoritmos de reconocimiento de emociones más complejos, donde se podían estimar de una forma más precisa las características de la voz. Actualmente, los investigadores se centran en crear aplicaciones de tiempo real eficientes, buscando combinaciones de distintas características de la señal de voz y/o clasificadores.

El reconocimiento automático de emociones se trata de una tarea muy compleja, ya que implica un amplio número de campos de investigación como pueden ser la lingüística, el análisis de voz, el aprendizaje automático o la psicología. De esta manera, para conseguir un progreso en este tipo de aplicaciones, se necesita avanzar en el desarrollo de cada uno de los campos implicados [1].

Se puede hablar de reconocimiento de emociones desde diferentes puntos de vista como pueden ser:

- **Psicológico**: Son muchas las teorías que se han ido formando sobre las emociones a lo largo de la historia del ser humano. En un principio, Descartes explicó que existe un conjunto primario de emociones que da lugar al estado emocional. Sin embargo, Darwin tenía la idea de que las emociones no se pueden separar de las costumbres más prácticas o duraderas, ya que han sido seleccionadas por la evolución debido a su valor para la supervivencia. Más tarde, William James hizo una descripción de las emociones relacionándolas con la percepción de la mente de condiciones fisiológicas que aparecen debido a un estímulo. De otra manera, Arnold mantuvo la idea de que las emociones tienen una valoración cognitiva que alerta al organismo en ciertas situaciones [3].

- **Biológico**: Se tiene la idea de que las respuestas emocionales y físicas se consideran patrones seleccionados para la supervivencia y que se han aprendido de una forma evolutiva. Se considera que las emociones tienen diferentes efectos en nuestro cuerpo como puede ser el aumento de la temperatura, la actividad muscular o el pulso. Estos cambios hacen que el estado emocional de una persona se manifieste con expresiones faciales y/o a través de la voz.

- **Lingüístico**: Es necesario realizar una clasificación de las emociones para etiquetar los diferentes estados emocionales, de modo que dichas etiquetas definan las distintas categorías que han de ser discriminadas por el sistema de reconocimiento automático de emociones [4].

2.2. Las emociones

2.2.1. La naturaleza de las emociones

Darwin relacionó la expresión sonora de las emociones al instinto, de modo que puede aparecer en una persona de una forma involuntaria. En ese momento empezó a considerarse la expresión emocional como el carácter del habla humana más transcultural y universal. El carácter transcultural del habla humana se ha investigado en diferentes estudios, donde se detectaron algunos elementos de influencia cultural que dejaron de lado la posición inicial de Darwin [5].

Scherer consideró que la expresión oral de las emociones se puede formar con un sistema analógico-vocal que se conecta con los mecanismos biológicos y fisiológicos del individuo. De esta forma se trabaja con estas hipótesis:

1. Cuando un individuo experimenta una emoción, en el cuerpo humano se producen unas alteraciones fisiológicas que producen cambios acústicos en la voz.
2. Cuando un hablante emocionado está hablando, el idioma modifica de alguna manera los rasgos acústicos producidos por las emociones.

Más tarde, Izard consideró que la emoción no es un fenómeno simple, sino que está constituida por muchos factores. En una de sus publicaciones explicó que una definición completa de emoción tiene que tener en cuenta los procesos que ocurren en el sistema nervioso y en el cerebro, y los modelos expresivos observables de la emoción [6].

Los términos emoción y estado de ánimo son conceptos diferentes. Las emociones surgen inesperadamente en respuesta a un determinado estímulo y tienen una corta duración de segundos o minutos. Sin embargo, los estados de ánimo pueden durar horas o incluso días y son ambiguos en su naturaleza. Las emociones se consideran como un estado cambiante, mientras que los estados de ánimo son más estables. Es imposible reconocer cuando una emoción se convierte en un estado de ánimo, por eso el término emoción se considera un término general que incluye al estado de ánimo. La personalidad de una persona se puede definir como el tono emocional característico que tiene una persona a lo largo del tiempo.

2.2.2. Clasificación de las emociones

Se considera que las emociones no son propias o características por si solas, sino que se dan como una combinación de algunas de ellas. Debido a esto, la clasificación de emociones es una tarea difícil y subjetiva. Sin embargo, los investigadores se refieren a emociones características o completas, ya que es la única forma de diferenciar unas emociones de otras.

A lo largo de la historia se han considerado diferentes clasificaciones de las emociones. Algunas de ellas son:

2.2.2.1. Emociones Positivas, Negativas o Neutras

Si la teoría se basa en el grado en el que las emociones afectan al comportamiento del individuo, se pueden clasificar las emociones en positivas, negativas o neutras. Se pueden definir emociones positivas y negativas según la intensidad y el grado de variabilidad. Los cambios de intensidad emocional pueden ser graduales o bruscos hacia lo positivo o hacia lo negativo. Una emoción puede representar una magnitud a lo largo de una línea continua, donde puede tomar valores positivos o negativos.

Las emociones positivas son aquellas que se experimentan cuando se alcanza algún objetivo, son emociones agradables. Por otra parte, las emociones negativas se producen cuando no se cumplen los objetivos marcados, cuando se siente frustración o se tiene una pérdida, es decir, se consideran emociones desagradables. Una tercera categoría son las emociones neutras, que son aquellas que no pertenecen a ninguna de las anteriores categorías, pero que sin embargo, comparte alguna característica con ambas. En la tabla 1 se muestra un ejemplo de cada tipo.

Emociones Positivas	Emociones negativas	Emociones Neutras
Felicidad	Miedo	Sorpresa
Alegría	Ansiedad	Esperanza
Amor	Ira	Compasión
Cariño	Hostilidad	
Humor	Tristeza	
	Vergüenza	
	Asco	

Tabla 1. Emociones negativas, positivas y neutras

En una emoción se pueden distinguir dos componentes:

- **Componente cualitativo**: La emoción se expresa a través de una palabra que describe la emoción y que define el signo positivo o negativo. Esas palabras pueden ser: “amistad”, “respeto”, “amor”, etc.
- **Componente cuantitativo**: Está definida por palabras que expresan magnitud independientemente del signo, como pueden ser “poco”, “mucho”, “nada”, “muy”, etc.

2.2.2.2. Emociones Primarias y Secundarias

Las emociones primarias son aquellas que surgen espontáneamente como respuesta a una situación, mientras que las emociones secundarias son una combinación de emociones básicas. Por ejemplo, si se reacciona con miedo (emoción primaria) ante una situación, las emociones secundarias que pueden surgir a partir de esta pueden ser la rabia o la ansiedad [E1].

Emociones Primarias:

- **Alegría**: Se reconoce por un incremento en la velocidad de locución y en la intensidad, así como en el tono medio y en su rango.
- **Tristeza**: Se conoce una voz triste cuando el tono medio es más bajo de lo habitual, además de presentar un rango de tono estrecho y una velocidad de locución lenta.
- **Enfado**: Se puede reconocer porque tiene un tono medio alto, un rango de tono amplio, además de una velocidad de locución rápida.
- **Disgusto/Odio**: Se reconoce porque tiene un tono medio bajo, un rango de tono amplio y su velocidad de locución es la más baja de todas.
- **Miedo**: Se caracteriza por tener el tono medio más elevado, el rango de tono mayor y una rápida velocidad de locución.

Emociones secundarias:

- **Ternura**: Presenta un alto nivel de tono.
- **Sorpresa**: Se caracteriza porque tiene un nivel de tono medio que está por encima del de la voz normal, sin embargo su velocidad de locución no se ve modificada, pero sí se detecta un rango de tono más amplio.
- **Ironía**: Se reconoce por una acentuación marcada y por tener una velocidad de locución baja.

- **Pena**: Se identifica por un tono medio bajo, tiene el rango de tono más estrecho que se conoce, y su velocidad de locución es baja, presentando un alto porcentaje de pausas.

Otras emociones secundarias como aburrimiento, queja, impaciencia, temor o satisfacción también han sido investigadas en diferentes estudios. En la tabla 2 se muestra una división de éste tipo de emociones.

Emociones Primarias	Emociones Secundarias
Alegría	Ternura
Tristeza	Sorpresa
Enfado	Ironía
Disgusto	Pena
Odio	Aburrimiento
Miedo	Queja
	Impaciencia
	Temor
	Satisfacción

Tabla 2. Emociones primarias y secundarias

2.2.2.3. Dimensiones del espacio semántico

Las dimensiones emocionales son una forma de representar las propiedades esenciales de las emociones de una manera simple.

Algunos autores como Suci, Tannembaum, Osgood, Davitz y Pereira han clasificado las emociones utilizando tres dimensiones del espacio semántico [7]:

- **Valencia, agrado o valoración:** Se representa en el eje horizontal y muestra el grado de positividad o negatividad que tiene la emoción en el individuo.
- **Actividad:** Se representa en el eje vertical, representa el grado de intensidad de la emoción.
- **Potencia o fuerza**: Se representa en el eje independiente. Se corresponde a las emociones que van desde la atención al rechazo, haciendo posible distinguir entre las emociones que inicia el individuo y las emociones que surgen del ambiente creado, por ejemplo en una conversación.

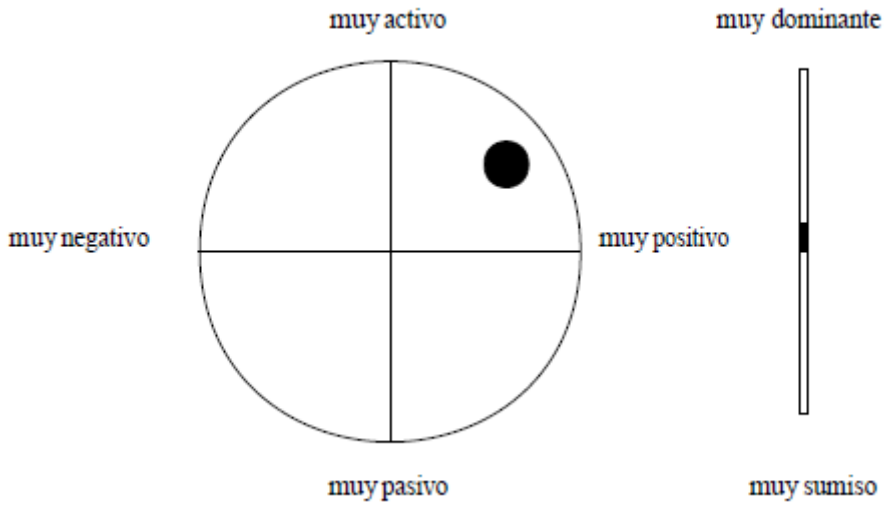

Figura 1. Dimensiones del espacio semántico, [7]

El investigador Marc Schroder sugiere que se representen todas las emociones en un espacio de tres dimensiones como el de la figura 1, con el fin de poder desarrollar mejores sistemas sintetizadores de voz expresiva.

En las descripciones de los estados emocionales en tres dimensiones se ha encontrado cierta correlación con algunas variables acústicas. Algunos investigadores como M. Schroder, M. Westerdijk, R. Cowie y S. Gielen [8] hablaron de estas correlaciones en algunas de sus publicaciones explicando qué se habían encontrado a partir de una base de datos de emociones.

2.2.2.4. Emociones Activas y Pasivas

Otros investigadores como Spinoza, dividieron las emociones en activas y pasivas [9]:

- **Activas**: Se pueden caracterizar por tener un tono y volumen alto, y por experimentar una rápida velocidad de locución.
- **Pasivas**: Se reconocen por un tono y volumen bajo, un timbre resonante y una velocidad de locución lenta.

En este estudio se trabaja con las emociones de la ira, el miedo, la alegría, la tristeza, el asco, el aburrimiento y la emoción neutra. Esta elección de emociones se debe a que son las que se representan en la base de datos que se toma de referencia para la extracción de características, y cuyo proceso de creación se explica en la sección 2.5.

2.2.3. Implicaciones Jurídicas

A la hora de hacer una sentencia legal, existen diferentes sectores donde las emociones están presentes y tienen un papel importante. A continuación se explican algunas de las influencias que pueden tener las emociones jurídicamente [10].

2.2.3.1. Valoración de las emociones en otras personas

En el sistema jurídico el reconocimiento de emociones a través de las características acústicas de la señal de voz tiene una gran utilidad. Gracias a esto, se puede evaluar la credibilidad de un sospechoso cuando está sometido a un interrogatorio. Los cambios en la señal de voz también pueden ayudar a un jurado a dar credibilidad a un testigo.

2.2.3.2. Emociones y Memoria

Las emociones tienen un papel muy importante en la memoria. Para añadir cierta fiabilidad a los testimonios de los testigos visuales o auditivos, se puede analizar su declaración para después hacer una valoración de su estado emocional. Los psicólogos cognitivos son capaces de distinguir entre formación, asociación y reconstrucción de la memoria, los cuales pueden estar afectados por las emociones.

2.2.3.3. Emociones y cultura

En las investigaciones forenses, pueden surgir problemas provenientes de las diferencias entre las emociones y la cultura. En los interrogatorios policiales, si los intérpretes no han sido entrenados correctamente, la interpretación de una lengua extranjera puede suponer un problema, ya que las traducciones literales suelen ser erróneas y no permiten entender las expresiones de una forma correcta.

2.2.3.4. Emociones y conocimiento legal

El sistema judicial está basado en normas morales, las cuales tienen integradas las emociones. Como ejemplo de ello se tiene que los crímenes se castigan, además de por su carácter intrínseco, por la actitud del culpable sobre la víctima. De esta manera se puede decir que el castigo se impone dependiendo de las emociones que expresa el culpable en el momento del acontecimiento.

En resumen, se puede afirmar que las emociones y la ley están directamente relacionadas.

2.3. El habla y el sistema de producción vocal

Desde el principio de los tiempos la voz ha sido el principal medio de comunicación entre los seres humanos. Independientemente del contenido del mensaje que se quiere transmitir, la voz contiene información sobre el hablante, sus emociones e incluso sobre su estado fisiológico. Desde hace ya varias décadas se están estudiando los mecanismos de producción del habla para poder crear sistemas automáticos de reconocimiento y síntesis de voz [11].

La voz es un sonido que producen las cuerdas vocales y por tanto se caracteriza por los siguientes elementos:

- **Intensidad**: Equivale al volumen del sonido. Cuando el aire sale de los pulmones, golpea en la glotis y produce vibraciones. Dependiendo de la amplitud de las vibraciones, se producirá una fuerza directamente proporcional. La intensidad transmite información sobre las emociones del hablante. Un volumen alto de voz se puede asociar al nerviosismo, tensión o agresividad; mientras que un volumen bajo puede asimilarse a una persona cansada o depresiva. Se mide en decibelios, dB, siendo el nivel medio en una conversación de 50 dB. En la figura 2 se pueden observar diferentes medidas de intensidad en ambientes cotidianos.

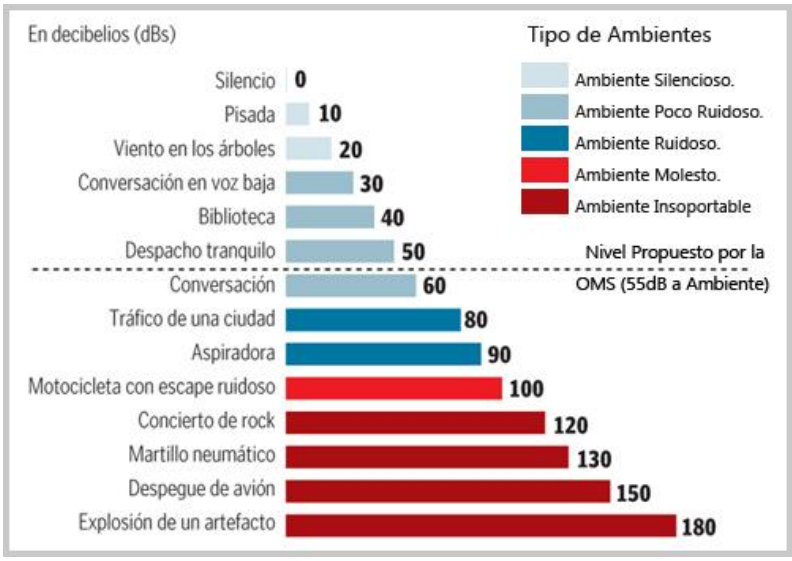

Figura 2. Intensidades en ambientes cotidianos [E2]

- **Tono**: Se relaciona con el número de vibraciones que se producen en la laringe del ser humano y que contiene una onda de sonido. Cuantas más vibraciones se producen, más aguda es la voz, es decir, el tono es más alto. Por el contrario, cuando se producen menos vibraciones, la voz es más grave y el tono más bajo. La unidad de medida del tono es el Hertzio o Hertz (Hz). No es fácil establecer la frontera entre un sonido grave y otro agudo, sin embargo, por consenso se suele establecer en torno a los 200Hz. Por debajo de ésta barrera oscilan las voces masculinas (80Hz a 200Hz), y por encima las voces femeninas (150Hz a 300Hz) [E3].

- **Timbre**: Permite diferenciar dos sonidos de igual intensidad y tono. Depende de la morfología de cada persona, ya que el aire sale de los pulmones y pasa a través de la laringe, los labios, los dientes y la lengua. El timbre es personal e independiente de cada persona. Puede aportar información real o imaginaria del locutor, como puede ser la apariencia física o la edad.

La variabilidad del habla es uno de los principales problemas que han encontrado los científicos. Sin embargo, se ha demostrado que con la voz pueden identificarse diferentes aspectos del estado físico y emocional del hablante como por ejemplo edad, sexo, apariencia, personalidad e incluso inteligencia. Estas características que son independientes de cada persona, contribuyen a la variabilidad del habla.

La voz se puede definir como el resultado de un proceso físico que se produce de forma voluntaria en el aparato fonador y que permite establecer una comunicación. El sistema nervioso central controla el sistema respiratorio y el digestivo, los cuales contienen órganos que intervienen en el proceso de producción vocal.

2.3.1. El aparato fonador

La voz se genera a través de una excitación en las cuerdas vocales que después se propaga a través de la faringe, la cavidad bucal y la cavidad nasal, las cuales se conocen como cavidades resonantes, ya que dependiendo de su forma se obtendrán unas características acústicas de la señal de voz distintas [12].

Las leyes de la acústica definen tres elementos que son necesarios para la producción del sonido: cuerpo vibrante, medio elástico que propague vibraciones y caja de resonancia que amplifique esas vibraciones para que el oído pueda percibirlas. El aparato fonador humano contiene todos esos elementos imprescindibles. Las cuerdas vocales situadas en la laringe representan el cuerpo vibrante, el medio de propagación es el aire que llega de los pulmones, y la caja de resonancia la componen la cavidad torácica, la faringe, las cavidades oral y nasal, y algunos elementos articulatorios como son los labios, los dientes, el paladar, el velo del paladar y la lengua.

La fonación tiene lugar durante la exhalación, cuando el aire que se encuentra en los pulmones sale y llega a la laringe a través de los músculos abdominales, los intercostales, el diafragma, los bronquios y la tráquea. El sonido se produce en la laringe cuando el aire choca con las cuerdas vocales.

Los distintos órganos que intervienen en el sistema de producción vocal, es decir, en la fonación, se pueden dividir en tres grupos:

2.3.1.1. Cavidades infraglóticas

Están formadas por los órganos que intervienen en la respiración (pulmones, bronquios y tráquea), siendo la fuente de energía del sistema de producción vocal.

Se pueden distinguir dos fases en el proceso de la respiración, las cuales se muestran en la Figura 3:

- **Inhalación**: En esta fase tiene lugar el impulso del soplo fonatorio, donde los pulmones se llenan de aire. El músculo encargado de esta tarea es el diafragma. También se conoce como inspiración.
- **Exhalación**: Se relaciona con el soplo fonatorio, donde el diafragma y la caja torácica se encargan de controlar y dosificar la salida del aire de los pulmones. También se conoce como espiración.

Figura 3. Cavidades infraglóticas [E4]

2.3.1.2. Cavidad laríngea

Lugar donde se encuentra el principal órgano fonador, la laringe. En esta cavidad, gracias a las cuerdas vocales, se modifica el flujo de aire que generan los pulmones y se convierte en una señal que excita las cavidades supraglóticas.

En la laringe se pueden diferenciar tres partes principales [E5]:

- **Supraglotis**: Parte superior de la laringe que incluye la epiglotis.
- **Glotis**: Parte media de la laringe, en la que se encuentran las cuerdas vocales.
- **Subglotis**: Es la parte inferior de la laringe, se encuentra entre la tráquea y las cuerdas vocales.

En la figura 4 se muestra una imagen de los órganos pertenecientes a ésta cavidad.

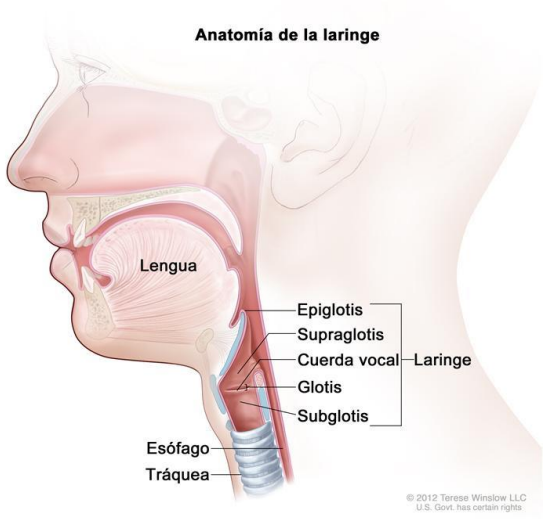

Figura 4. Cavidad laríngea [E5]

2.3.1.3. Cavidades supraglóticas

Las cavidades supraglóticas o tracto vocal, están formadas por la faringe, la cavidad oral y la cavidad nasal. Su función es modificar el flujo de aire que proviene de la laringe, para generar una señal acústica a la salida de la nariz y la boca.

En la figura 5 se pueden ver los principales órganos del aparato fonador, y en particular del tracto vocal, como son los labios, los dientes, el paladar, la lengua, etc.

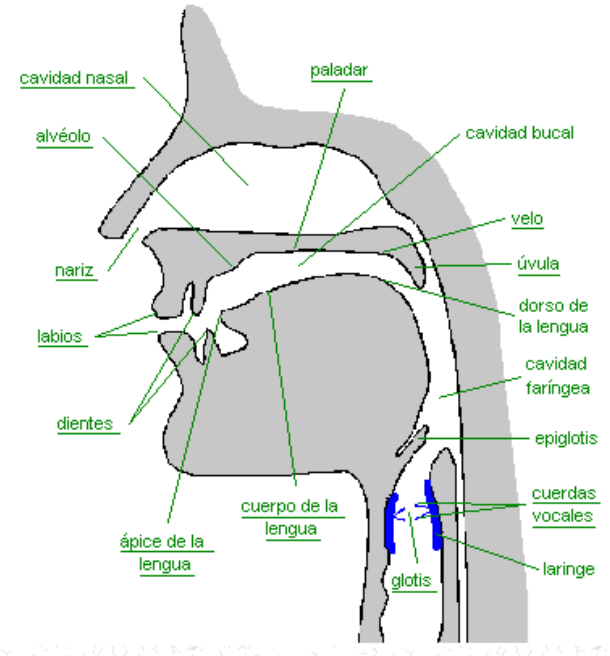

Figura 5. Cavidades supraglóticas [E6]

2.3.2. Modelo de producción vocal

El aire sale de los pulmones pasando por la tráquea y la glotis, con una presión diferente dependiendo del sonido que se quiera generar. La glotis es el mecanismo que separa las cuerdas vocales. Éstas se mantienen abiertas en el momento de la respiración y se van estrechando según los sonidos que se quieran producir. La velocidad con la que se abren y cierran las cuerdas vocales es lo que se denomina frecuencia fundamental o Pitch. Una vez que el aire sale de la glotis, llega al tracto vocal. Cada persona tiene un tracto vocal diferente, es decir, dependiendo de su forma, la posición de los órganos articuladores y resonadores variará, y los sonidos producidos serán distintos. Las resonancias que se producen tienen concentrada su energía alrededor de algunas frecuencias del espectro, lo que se conoce como formantes (se hablará de ellos en el capítulo 3).

En la figura 6 se representa un ejemplo del mecanismo de producción vocal.

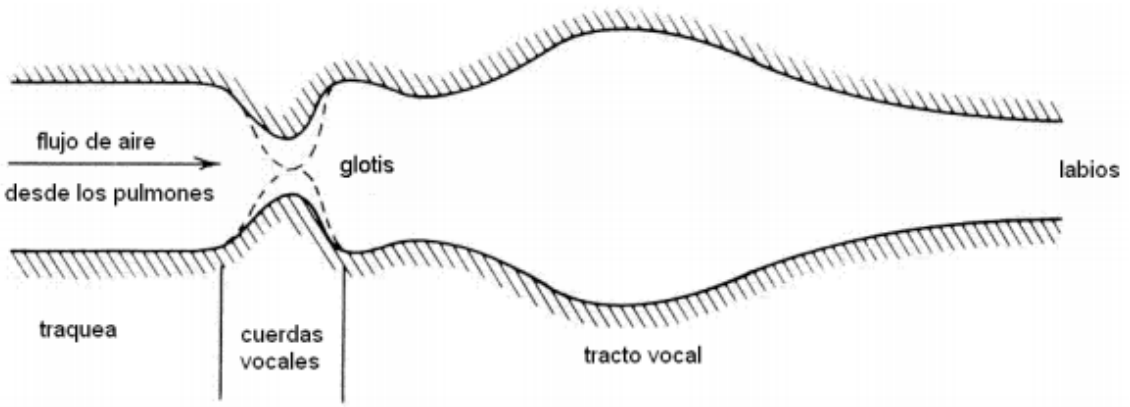

Figura 6. Mecanismo de producción vocal [E7]

2.3.3. Análisis de predicción lineal

El análisis de predicción lineal se trata de una técnica muy utilizada para la caracterización de la voz, basada en el modelo general de producción vocal conocido como "Modelo de tubos", representado en la figura 7. El tracto vocal se puede modelar como una concatenación de tubos acústicos con un diámetro diferente. Este sistema da lugar a un modelo lineal no estacionario, donde las secciones de los tubos van cambiando dependiendo de los fonemas que se emiten [E7].

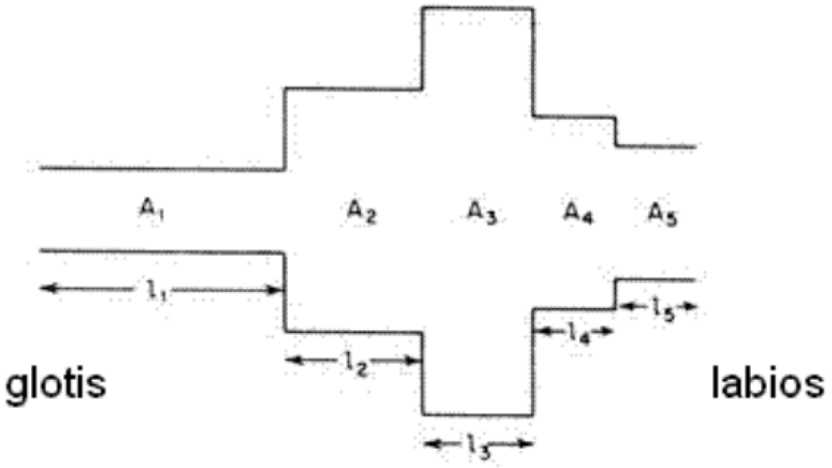

Figura 7. Sistema de producción vocal, modelo de tubos [E7]

En la figura 8 se representa el proceso de producción vocal, definido por una entrada (aire), un filtro (tracto vocal) y una salida (voz). Los parámetros del tracto vocal pueden variar en el tiempo dependiendo de la acción que se realice al pronunciar una palabra. Se pueden diferenciar dos tipos de señales de entrada en el filtro: sonora y no sonora. Cuando la señal es sonora, la excitación es un tren de impulsos de frecuencia determinada, sin embargo, si la señal no es sonora, la excitación será un ruido aleatorio. El funcionamiento de la glotis se modela con la combinación de ambas señales.

La señal de entrada pasa por el conducto vocal, también conocido como tracto vocal, donde se amplifica y se ve modificada por sus frecuencias de resonancia (formantes). A la salida se añaden los efectos de radiación que producen los labios.

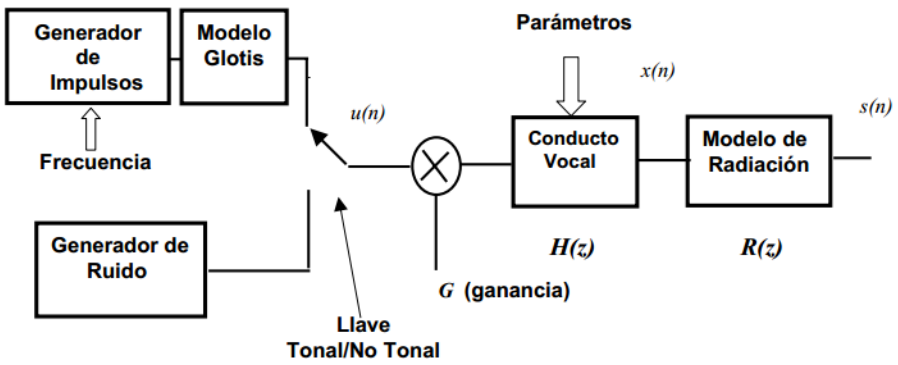

Figura 8. Sistema de producción vocal, [E7]

En la figura 8 se pueden observar tres elementos básicos:

- **Tracto vocal o conducto vocal**: Se representa a través de una función todo polos.

$$H(z) = \frac{G}{1 + \sum_{i=1}^{N} a_i z^{-i}}$$

Ecuación 1. Tracto vocal

- **Modelo de radiación**: Se utiliza un filtro paso alto para introducir el efecto de la radiación de los labios en el modelo digital.

$$R(z) = (1 - \alpha z^{-1})$$

Ecuación 2. Modelo de radiacción

Dependiendo del tipo de voz que se produce se utiliza una **fuente de excitación** distinta. Para voz sonora, se genera un tren de impulsos que excita un sistema lineal, mientras que para la voz sorda se utiliza un generador de ruido aleatorio. Esto se combina con el modelo de la Glotis $G(z)$.

Para obtener la función de transferencia global se necesitan combinar los sistemas anteriores:

$$F(z) = G(z)H(z)R(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{i=1}^{N} a_i z^{-i}}$$

Ecuación 3. Función de transferencia simplificada

La ecuación 3 corresponde a considerar únicamente H(z), es decir, se trata de una función de transferencia simplificada donde no se tienen en cuenta los modelos de glotis y de radiación.

Por tanto, la señal de voz se puede modelar como:

$$s[n] = G\, u[n] - \sum_{k=1}^{P} a_k\, s[n-k]$$

Ecuación 4. Señal de voz modelada

En la ecuación 4, s[n] se corresponde con la señal de voz y u[n] es la excitación.

$$u[n] = \begin{cases} Ruido\ blanco\ y\ gaussiano & Voz\ sonora \\ \frac{1}{N_0}\sum_{r}\delta[n - rN_0] & Voz\ sorda \end{cases}$$

Ecuación 5. Tipos de excitación de la señal de voz

Donde N_0 es el periodo fundamental (inverso de la frecuencia fundamental o F_0).

Con todo lo anterior, obtenemos un sistema del tipo:

$$E = \sum_{n=-\infty}^{\infty}\left(s[n] + \sum_{k=1}^{P} a_k\, s[n-k]\right)^2$$

Ecuación 6. Sistema de producción vocal

Derivando e igualando a cero la ecuación 6 se obtiene un sistema de ecuaciones con el que es posible calcular los parámetros a_k que minimicen dicha ecuación. Dichos coeficientes se denominan coeficientes de predicción lineal, o por sus siglas en inglés, LPC ("Linear Prediction Coefficients").

2.3.4. Efecto de las emociones en el habla

El efecto de las emociones en el habla ha sido investigado por lingüistas que han estudiado los efectos léxicos y prosódicos, por psicólogos, y por profesionales de la acústica que han analizado la señal de voz.

Se tienen que tener en cuenta dos factores fundamentales para implementar de una manera satisfactoria un sistema automático de reconocimiento de emociones en el habla: hay que saber distinguir las características emocionales de la voz y se necesitan conocimientos sobre cómo usar los métodos convencionales de procesado de voz para poder describir dichas características.

Los principales componentes de la voz que pueden ser analizados y que expresan emociones, son la frecuencia fundamental (F_0 o pitch), la calidad de voz, la duración y la forma del tracto vocal [13]. Se hablará de estas características con profundidad en el capítulo 3.

2.4. Sistemas de reconocimiento de emociones

Los primeros estudios acerca de sistemas de reconocimiento automático de emociones a través de las expresiones faciales se realizaron en el año 1992 [14], mientras que el reconocimiento automático de emociones a partir del habla empezó a estudiarse a finales de la década de los años 90.

Los primeros estudios relacionados con el reconocimiento de emociones no buscaron obtener un sistema de reconocimiento eficiente, sino que se centraron en buscar propiedades acústicas del habla que estuvieran relacionadas con las emociones, como por ejemplo, que la felicidad tiende a tener una media de la frecuencia fundamental F_0, o pitch, más alta que en frases neutras [13].

Más adelante, se empezó a creer que la computación afectiva tenía un importante potencial industrial [15], y esto impulsó la investigación hacia la búsqueda de buenos rendimientos en el reconocimiento automático de las emociones en el habla.

En la actualidad, se están realizando estudios con un gran número de clasificadores y utilizando características muy diversas; mientras que en estudios anteriores sólo utilizaban un número reducido de técnicas de aprendizaje y/o características acústicas [16]. Además en algunos estudios se utilizaban bases de datos con muy pocas muestras [17]. Sólo en [18] se han utilizado más características que las pertenecientes a un conjunto estándar como pueden ser media, mínimo, máximo, varianza de las distribuciones de intensidad y tono, o las duraciones de los fonemas.

Otro inconveniente que presentan algunos estudios, es que las bases de datos de voz con las que se trabajan han sido grabadas por personas que tratan de leer textos de forma emocional (por ejemplo, periodísticos). En este sentido, últimamente se han realizado dos investigaciones [16] que han tratado de construir sistemas de reconocimiento automático del habla espontánea.

La metodología que se usa en este trabajo está basada en el estudio [2], donde el objetivo es usar unas herramientas parecidas, y con la misma base de datos propuesta en dicho estudio, realizar experimentos donde se extraigan diferentes tipos de características para alcanzar unos resultados similares o incluso mejorarlos.

2.5. Bases de datos

2.5.1. Introducción

Para los diversos estudios realizados sobre el reconocimiento de emociones se han utilizado diferentes muestras de voz provenientes, en su mayoría, de unas bases de datos. Actualmente existen un gran número de bases de datos que contienen grabaciones de voz realizadas tanto en estudios, como en condiciones más realistas. En algunos casos, las muestras han sido grabadas por actores profesionales, pero también existen otras muestras espontáneas del hablante, incluso la interacción de personas adultas con niños. Cuanto mayor sea la diversidad y magnitud de la base de datos, más realistas serán los resultados obtenidos.

En la tabla 3 se muestran diferentes bases de datos de habla emocional utilizadas para el reconocimiento y síntesis de emociones. Se incluyen distintas características de cada base de datos como puede ser el idioma en el que se han grabado, las emociones expresadas, características de los locutores y si los datos han sido simulados o son naturales.

Referencia	Lenguaje	Locutores	Tipo
Abelin and Allwood (2000)	Sueco	1 Nativo	Simulada
Alpert et al. (2001)	Inglés	22 enfermos y 19 sanos	Natural
Alter et al. (2000)	Alemán	1 Mujer	Simulada
Ambrus (2000), Interface	Inglés, Eslovenio	8 Actores	Simulada
Amir et al. (2000)	Hebreo	40 estudiantes	Natural
Ang et al. (2002)	Inglés		Natural
Banse and Scherer (1996)	Alemán	12 Actores	Simulada
Batliner et al. (2004)	Alemán, Inglés	51 Niños	Forzada
Bulut et al. (2002)	Inglés	1 Actriz	Simulada
Burkhardt and Sendlmeier (2000)	Alemán	10 Actores	Simulada
Caldognetto et al. (2004)	Italiano	1 Nativo	Simulada
Choukri (2003), Groningen	Holandés	238 Nativos	Simulada
Chuang and Wu (2002)	Chino	2 Actores	Simulada
Clavel et al. (2004)	Inglés	18 Personas TV	Simulada
Cole (2005), Kids' Speech	Inglés	780 Niños	Natural
Cowie and Douglas-Cowie (1996)	Inglés	40 Nativos	Natural
Douglas-Cowie et al. (2003)	Inglés	125 Personas TV	Seminatural
Edgington (1997)	Inglés	1 Actor	Simulada
Engberg and Hansen (1996),	Danés	4 Actores	Simulada

Fernandez and Picard (2003)	Inglés	4 Conductores	Natural
Fischer (1999), Verbmobil	Alemán	58 Nativos	Natural
France et al. (2000)	Inglés	70 Enfermos, 40 Sanos	Natural
Gonzalez (1999)	Inglés, Español		Forzada
Hansen (1996), SUSAS	Inglés	32	Natural, Simulada
Heuft et al. (1996)	Alemán	3 Nativos	Simulada, Forzada
Iida et al. (2000), ESC	Japonés	2 Nativos	Simulada
Iriondo et al. (2000)	Español	8 Actores	Simulada
Kawanami et al. (2003)	Japonés	2 Actores	Simulada
Lee and Narayanan (2005)	Inglés	Desconocido	Natural
Liberman (2005), Emotional Prosody	Inglés	Actores	Simulada
Linnankoski et al. (2005)	Inglés	13 Nativos	Forzada
Lloyd (1999)	Inglés	1 Nativo	Simulada
Makarova and Petrushin (2002),	Ruso	61 Nativos	Simulada
Martins et al. (1998), BDFALA	Portugués	10 Nativos	Simulada
McMahon et al. (2003), ORESTEIA	Inglés	29 Nativos	Forzada
Montanari et al. (2004)	Inglés	15 Niños	Natural
Montero et al. (1999), SES	Español	1 Actor	Simulada
Mozziconacci and Hermes (1997)	Holandés	3 Nativos	Simulada
Niimi et al. (2001)	Japonés	1 Hombre	Simulada
Nordstrand et al. (2004)	Sueco	1 Nativo	Simulada
Nwe et al. (2003)	Chino	12 Nativos	Simulada
Pereira (2000)	Inglés	2 Actores	Simulada
Petrushin (1999)	Inglés	30 Nativos	Simulada, Natural
Polzin and Waibel (2000)	Inglés	Desconocido	Simulada
Polzin and Waibel (1998)	Inglés	5 estudiantes de teatro	Simulada
Rahurkar and Hansen (2002), SOQ	Inglés	6 Soldados	Natural
Scherer (2000b) Lost Luggage	Varios	109 Pasajeros	Natural
Scherer (2000a)	Alemán	12 Actores	Simulada
Scherer et al. (2002)	Inglés, Alemán	100 Nativos	Natural
Schiel et al. (2002), SmartKom	Alemán	45 Nativos	Natural
Schröder and Grice (2003)	Alemán	1 Hombre	Simulada
Schröder (2000)	Alemán	6 Nativos	Simulada
Slaney and McRoberts (2003)	Inglés	12 Nativos	Natural
Stibbard (2000), Leeds	Inglés	Desconocido	Natural, Forzada

Tato (2002), AIBO	Alemán	14 Nativos	Forzada
Tolkmitt and Scherer (1986)	Alemán	60 Nativos	Forzada
Wendt and Scheich (2002)	Alemán	2 Actores	Simulada
Yildirim et al. (2004)	Inglés	1 Actriz	Simulada
Yu et al. (2001)	Chino	Nativos TV	Simulada
Yuan (2002)	Chino	9 Nativos	Forzada

Tabla 3. Bases de datos de emociones, adaptada de [19]

A continuación se describe la base de datos utilizada en este proyecto.

2.5.2. Berlin Data Base

Esta base de datos alemana de expresión emocional actuada, contiene diez oraciones pertenecientes a seis emociones diferentes realizadas por diez actores (ira, miedo, alegría, tristeza, aburrimiento, asco y neutra) [20].

2.5.2.1. Emociones reales o actuadas

Al escoger esta base de datos, se ha sido consciente de que las emociones expuestas han sido representadas por actores, lo cual puede hacer perder credibilidad al estudio, puesto que las verdaderas emociones son muy difíciles de grabar en la vida real. Además, hay señales emocionales físicas que no pueden ser imitadas conscientemente.

Para las necesidades de este proyecto debían cumplirse los siguientes puntos:

- Un número razonable de personas debe interpretar todas las emociones para ofrecer una generalización sobre el grupo a analizar.
- Todos los hablantes deben pronunciar el mismo contenido verbal con el fin de permitir la comparabilidad entre las emociones y los hablantes.
- Las grabaciones deben ser de alta calidad de audio, lo que minimiza el ruido de fondo. De lo contrario, las mediciones espectrales pueden estar contaminadas por el ruido. En este sentido, las grabaciones deben realizarse en una cámara anecoica, para evitar tanto el eco y otras reflexiones de la onda sonora, como el ruido de fondo.

También se ha tenido en cuenta el problema de que cada persona reacciona de manera distinta a las situaciones emocionales. Se confía en la capacidad de

interpretación de los artistas por recordar una situación en la que la emoción deseada se ha sentido con fuerza, lo que se conoce como el método de Stanislavski [21].

2.5.2.2. Elección de emociones

Con el fin de poder comparar los resultados obtenidos con estudios anteriores del mismo ámbito de investigación, se utilizaron las mismas etiquetas emocionales, las cuales se muestran en la tabla 4.

Emoción	Traducción
Neutral	Neutra
Ärger	Ira
Angst	Miedo
Freude	Alegría
Trauer	Tristeza
Ekel	Asco
Langeweile	Aburrimiento

Tabla 4. Emociones de "Berlin Data Base"

2.5.2.3. Elección de actores

Teniendo en cuenta que los actores aprenden a expresar sus emociones de una manera bastante exagerada, se consideró que actores profesionales no serían la mejor opción para llevar a cabo las expresiones emocionales naturales. De esta manera, se decidió buscar artistas por medio de un anuncio en el periódico. Unas 40 personas respondieron y fueron invitadas a una sesión de preselección. Tenían que interpretar una expresión que representara a cada una de las emociones establecidas, las cuales se grabaron en una oficina con un micrófono directamente conectado a un disco duro. De estas 40 sesiones, tres oyentes expertos seleccionaron a 10 personas, en representación de los dos sexos por igual, a juzgar por la naturalidad y carácter reconocible de la actuación. Curiosamente, todos menos uno de los elegidos habían realizado algún curso de interpretación.

En la tabla 5 se muestran las características de los actores elegidos.

Nº identificador actor	Persona	Edad
3	Hombre	31
8	Mujer	34
9	Mujer	21
10	Hombre	32
11	Hombre	26
12	Hombre	30
13	Mujer	32
14	Mujer	35
15	Hombre	25
16	Mujer	31

Tabla 5. Actores de "Berlin Data Base"

2.5.2.4. Materiales de texto

Es importante que todas las frases escogidas interpreten las emociones objeto de examen y que no contengan ningún contenido emocional. Existen dos tipos de materiales que suelen cumplir estos objetivos:

- Material de texto sin sentido, como por ejemplo una serie desordenada de cifras o letras, o palabras de fantasía.
- Frases normales que podrían ser utilizadas en la vida cotidiana.

Los materiales sin sentido garantizan que sean emocionalmente neutros. Sin embargo, existe el inconveniente de que a los actores les resulta difícil imaginar una situación emocional con este tipo de frases, y así poder interpretar de forma espontánea esa situación.

En comparación con los poemas y frases sin sentido, el uso de la comunicación diaria ha demostrado ser mejor, ya que esta es la forma natural de expresión bajo la excitación emocional. Por otra parte, los actores pueden hablar desde el recuerdo de alguna situación, no hay necesidad de un proceso de memorización o la lectura de un texto. En la construcción de la base de datos, se dio prioridad a la naturalidad del material de discurso y, por tanto se utilizaron frases cotidianas como enunciados de prueba. Se construyeron un total de diez frases; cinco frases simples y cinco compuestas de dos oraciones simples.

Para poder llevar a cabo el análisis de formantes, las frases de prueba tenían que contener tantas vocales como fuera posible.

Las frases usadas se muestran en la tabla 6.

Nº Identificación	Frase	Traducción
a01	Der Lappen liegt auf dem Eisschrank	El trapo cuelga de la nevera
a02	Das will sie am Mittwoch abgeben	Ella va a entregarlo el miércoles
a04	Heute Abend könnte ich es ihm sagen	Esta noche puedo contárselo
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück	La hoja de papel negro está ahí arriba al lado de la pieza de madera
a07	In sieben Stunden wird es soweit sein	En siete horas habrá llegado el momento
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	¿Cuáles son las bolsas que estaban allí debajo de la mesa?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter	Acaban de llegar arriba y ahora bajan de nuevo
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht	Los fines de semana siempre voy a casa y visito a Agnes
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen	Sólo quiero quitarme esto y después ir a tomar una copa con Karl
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen	Estará en el sitio donde siempre lo ponemos

Tabla 6. Expresiones de "Berlin Data Base"

2.5.2.5. Grabaciones

Para lograr una alta calidad en el audio de las grabaciones, estas se realizaron en la cámara anecoica de la Universidad Técnica de Berlín. Las grabaciones fueron tomadas con una frecuencia de muestreo de 48 kHz y más tarde se submuestrearon a 16 kHz.

Los actores se colocaban como muestra la figura 9, de pie delante del micrófono para poder utilizar el lenguaje corporal si lo deseaban. Era necesario que hablaran en la dirección del micrófono y a una distancia de unos 30 cm.

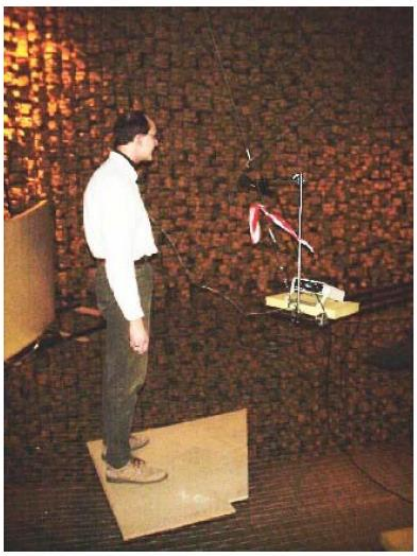

Figura 9. Posición del actor frente al micrófono, [20]

Hubo una sola sesión de grabación la cual tuvo una duración de alrededor de dos horas con cada uno de los actores. Estas se realizaron bajo la supervisión de tres profesionales de la fonética, dando dos de ellos instrucciones y comentarios, y el otro supervisando los equipos de grabación. El texto de cada una de las expresiones se enseñaba mediante un papel al actor para evitar un estilo de entonación lectora. Después, cada actor escuchó una breve caracterización de cada emoción (por ejemplo, la felicidad después de ganar una gran cantidad de dinero en la lotería o la tristeza causada por la pérdida de un buen amigo o familiar) y les dieron tiempo para ponerse en el papel de esa emoción específica. A los actores se les pedía que recordaran una situación real de su pasado donde habían sentido dicha emoción. De esta manera se obtuvieron grabaciones de actores que re-experimentan las emociones y que desarrollaron los mismos efectos fisiológicos que en la situación real.

Cada actor podía grabar la frase de cada emoción tantas veces como quisiera, además fueron instruidos para no gritar al expresar la ira o evitar susurrar al tiempo que expresaban la tristeza. Esto era necesario con el fin de obtener datos analizables en relación con la calidad de voz.

Después de todo, existían algunos problemas como:

- Debido a que los actores se encontraban de pie frente al micrófono, al hacer gesticulaciones la distancia entre la boca y el micrófono era variable, y por tanto el análisis de la energía de la señal no podía ser fiable.
- El nivel de grabación tuvo que ser ajustado, ya que en algunas emociones la voz se interpreta más alta, como puede ser la ira, y sin embargo en otras, el nivel del habla es más bajo, como por ejemplo en la tristeza.
- Otro problema es aplicable a la curva de entonación: cada actor elegía diferentes palabras para la realización del acento en cada frase.

2.5.2.6. Evaluación de los datos

Para asegurar la calidad emocional y la naturalidad de las expresiones se llevó a cabo una prueba de percepción en la que participaron 20 personas. A cada una de ellas se le dejó un ordenador y se le fueron mostrando las frases en orden aleatorio. Únicamente se les permitió escuchar cada muestra una vez, y de esta forma evaluar en qué estado emocional situaban al hablante y lo convincente que era la actuación.

En la figura 10 se muestran las medias de la tasa de reconocimiento de cada emoción. Las líneas de conexión entre las barras muestran diferencias significativas entre las emociones ($p < 0,05$). Las frases con una tasa de reconocimiento mayor al 80% y una naturalidad mayor al 60% se eligieron para su posterior análisis. En total, se descartaron unas 500 interpretaciones de las 800 que fueron grabadas.

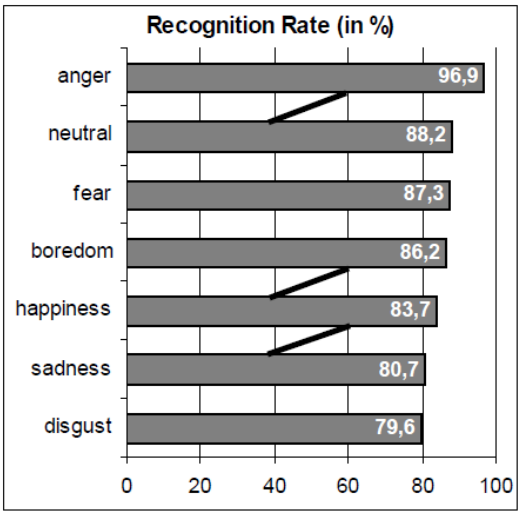

Figura 10. Tasa de reconocimiento de las emociones de "Berlin Data Base", [20]

Más tarde se realizaron dos pruebas más de percepción. En una de ellas se pidió a los individuos que calificaran la intensidad de la emoción que se mostraba en cada frase. Sin embargo, en la segunda prueba, los sujetos tenían que señalar la sílaba tónica de cada expresión. Esta prueba fue la única en la que sólo las personas que entendían de fonética podían participar, ya que la mayoría de las personas no estaban cualificadas para poder diferenciar las sílabas tónicas. En ambas pruebas se daba la opción de poder escuchar las expresiones tantas veces como se quisiera antes de dar su calificación.

Los resultados mostraron que la mayoría de las emociones empiezan en un tono moderado y terminan en otro más fuerte. De esta forma, la intensidad emocional se utilizó como variable de control en los análisis estadísticos.

2.5.2.7. Etiquetado de los datos

Para el etiquetado de los datos se crearon dos ficheros de etiquetas en formato ASCII por cada expresión. El primer fichero de etiquetas contiene una transcripción fonética que se basa en un juicio auditivo apoyado por un análisis visual realizado por un oscilograma y por un espectrograma. Para la transcripción se utilizó el alfabeto fonético SAMPA [22]. Algunas características emocionales de la voz y la manera de hablar se marcaron con caracterizaciones adicionales. El segundo archivo de etiquetas contiene una segmentación en sílabas y marcas con cuatro niveles diferentes de acentuación.

2.5.2.8. Presentación de los datos

Para la presentación de esta base de datos de expresión emocional se desarrolló una interfaz web donde se encuentra toda la información disponible y que se puede acceder a través de Internet [E8].

Ahí se puede seleccionar cada uno de los enunciados de la base de datos y escucharlos. La elección se puede hacer según el hablante, el texto hablado o la emoción expresada. Además está disponible la descarga de la base de datos donde se encuentran las muestras de voz grabadas y etiquetadas.

Por cada expresión se puede representar la información de las sílabas, la duración de las curvas de entonación, los histogramas de frecuencia fundamental, la vibración de las cuerdas vocales, la energía y las curvas de volumen. También se muestran los resultados de las pruebas de evaluación. Como característica especial se pueden escuchar y descargar diferentes versiones del enunciado inicial.

2.6. Métodos de clasificación

2.6.1. Introducción

La clasificación es el proceso de asignar objetos a un conjunto de categorías o clases. Teniendo en cuenta el caso que se está estudiando, las categorías se determinan por las emociones que se quieren reconocer a partir de una serie de características medidas sobre un conjunto de individuos.

Las técnicas de clasificación se pueden dividir en métodos supervisados y en métodos no supervisados. Se considera un método supervisado cuando el clasificador se entrena a partir de una base de datos de entrenamiento etiquetada, mientras que se considera un método no supervisado cuando no hace falta disponer de datos etiquetados para el entrenamiento del clasificador [23].

Existen diferentes técnicas de clasificación supervisada que se utilizan en el reconocimiento de emociones, como son las redes neuronales artificiales ("Artificial Neural Networks", ANN), las máquinas de vector soporte ("Support Vector Machines", SVM), los árboles de decisión, los modelos ocultos de Markov ("Hidden Markov Models", HMM), las redes bayesianas, los algoritmos de votación Bagging y Boosting, y los modelos de mezclas de gaussianas ("Gaussian Mixture Models", GMM).

A continuación se explican las dos técnicas más utilizadas en estos tipos de sistemas de reconocimiento de emociones.

2.6.2. Modelos ocultos de Markov

Los modelos ocultos de Markov es la técnica de reconocimiento automático del habla más utilizada. Se considera que un HMM es una máquina de estados donde todos se encuentran conectados entre sí a través de unas probabilidades de transición a_{ij}. En la figura 11 se puede apreciar este tipo de modelo donde cada estado crea un vector de observación teniendo en cuenta una determinada función de densidad de probabilidad [24].

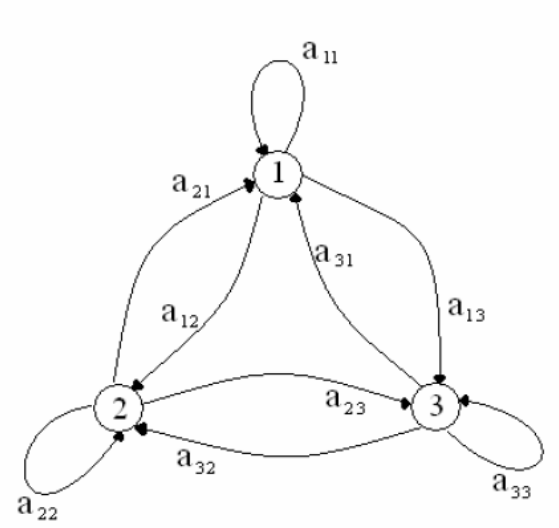

Figura 11. Modelo oculto de Markov, [24]

La salida del modelo se determina según el número de estados recorridos, sin embargo, no se puede observar la secuencia de estados por los que se ha pasado, de ahí que se denominen modelos ocultos. La tarea fundamental consiste en identificar los parámetros ocultos a partir de los parámetros observados. Aunque los estados no son visibles para el observador, las variables pertenecientes a cada estado sí que lo son.

Un HMM está compuesto por [25]:

- N estados
- π: matriz de probabilidades iniciales
- A: matriz de transiciones de estados
- B: función de densidad de probabilidad asociada a cada estado.

En la figura 11 representada anteriormente, se pueden observar tres estados, por tanto N=3. Se define q_t = N a la probabilidad de que el modelo se encuentre en el estado N, en el instante t. En este caso, la matriz de transiciones sería:

$$A = \begin{Bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Bmatrix}$$

Ecuación 7. Matriz de transiciones para 3 estados

Si se generaliza para N estados, la matriz de transiciones A tendría la siguiente forma:

$$A = \{a_{ij}\}, 1 \leq i, j \leq N$$

Ecuación 8. Matriz de transiciones para N estados

Donde a_{ij} representa la probabilidad de que estando en un estado i en un cierto instante de tiempo, se pase al estado j en el siguiente instante:

$$a_{ij} = P[q_{t+1} = j \mid q_t = i]$$

Ecuación 9. Probabilidad de un estado

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \leq i \leq N$$

Ecuación 10. Criterio de la probabilidad

Si se quiere representar la probabilidad de observar o_t (observación en el instante t) en el estado j, se define una función de densidad de probabilidad que tiene la siguiente forma:

$$b_j(o_t) = P[o_t \mid q_t = j]$$

Ecuación 11. Probabilidad de observación en el instante t

Otro elemento que es necesario definir en el modelo, es el vector de inicialización llamado π.

$$\pi = \{\pi_i\}, 1 \leq i \leq N$$

$$\pi_i = P[q_i = i]$$

Ecuación 12. Vector de inicialización

Donde π_i es la probabilidad de comenzar la secuencia de observación en cada uno de los estados i.

En resumen, un modelo oculto de Markov se puede expresar como:

$$\lambda = (A, B, \pi)$$

Ecuación 13. Ecuación HMM

2.6.3. Máquinas de vector soporte

Las máquinas de vector soporte ("Support Vector Machines", SVMs) son una técnica de clasificación basada en la teoría estadística del aprendizaje, introducida por V. Vapnik [26]. SVM se puede definir como una clase específica de algoritmos preparados para el entrenamiento de una máquina de aprendizaje lineal en un espacio que contiene una función núcleo, teniendo en cuenta unas reglas de generalización.

SVM es una técnica que se ha ido desarrollando en las dos últimas décadas en el campo de la clasificación automática de patrones. Se trata de un modo de clasificación que ha conseguido solventar algunos de los problemas de las redes neuronales y de los modelos ocultos de Markov gracias a su gran capacidad de discriminación. Una de las limitaciones que presenta esta técnica es que los vectores de entrada tienen que ser de longitud fija.

Existen dos ideas fundamentales para construir un clasificador SVM:

- Primero hay que mapear los puntos de entrada a un espacio de características que tenga una dimensión mayor, es decir, transformar el espacio de entrada en un espacio de alta dimensión, figura 12.

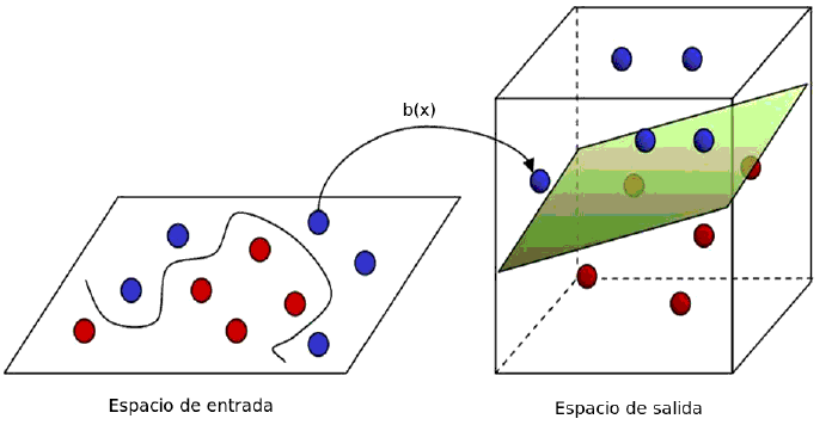

Figura 12. SVM, Mapeo de punto a un espacio de mayor dimensión, [27]

- En ese espacio, hay que localizar un hiperplano que separe las clases, como se puede apreciar en la figura 13.

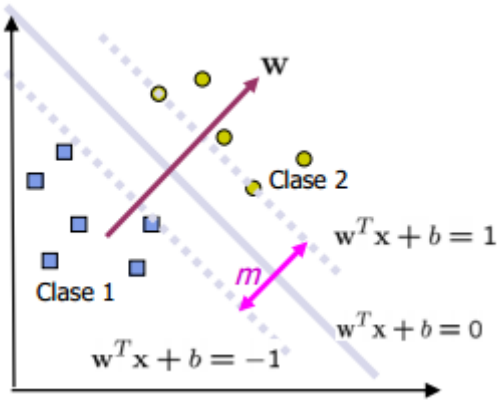

Figura 13. SVM, Hiperplano separador de clases, [27]

El objetivo de este tipo de clasificadores es maximizar el margen "m" entre clases. Este tipo de cálculos es un problema de programación cuadrática (QP) que puede ser resuelto introduciendo multiplicadores de Lagrange [27].

El clasificador SVM encuentra un hiperplano óptimo que se crea a partir de las muestras pertenecientes al conjunto de entrenamiento combinadas con algunos puntos de entrada llamados vectores soporte. Cuanto más pequeño sea el número del vector soporte, el hiperplano se encontrará situado en el punto más óptimo.

Cuando los datos disponibles no son linealmente separables a través de un hiperplano, como se muestra en la figura 14, se utilizan unas funciones llamadas "Kernel". El algoritmo que tienen este tipo de funciones transforma los puntos de entrada en puntos de mayor dimensión antes de la clasificación.

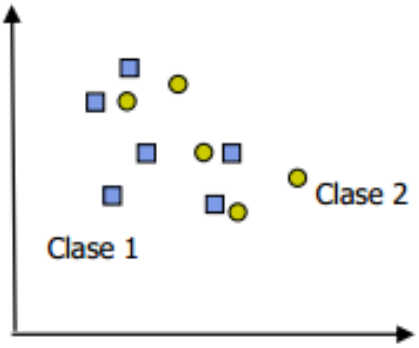

Figura 14. SVM, datos linealmente no separables, [27]

Las funciones Kernel más utilizadas son:

- **Kernel lineal**:

$$k(x_i, x_j) = x_i . x_j$$

Ecuación 14. Kernel lineal

- **RBF kernel**: Con esta función se consiguen buenos resultados en la resolución de este tipo de problemas, además de proporcionar una buena generalización.

$$k(x_i, x_j) = e^{-\left(\gamma |x_i . x_j|^2\right)}$$

Ecuación 15. RBF Kernel

γ se define como la distancia entre el hiperplano y las muestras que se encuentran más cerca de éste, por tanto, para minimizar el error, se puede elegir un margen menor de γ.

Por tanto, se puede definir la regla de clasificación para un clasificador binario como [58]:

$$f(x) = b + \sum_i \alpha_i \, K(x, x_i)$$

Ecuación 16. Clasificador binario

Donde b y α_i son parámetros que el clasificador ha aprendido durante el proceso de entrenamiento. $K(x, x_i)$ es el valor que tiene la función kernel en los puntos x y x_i.

En este proyecto de reconocimiento de emociones se trabaja con 6 emociones, por tanto habría que generalizar el clasificador binario al caso multiclase. Existen diferentes algoritmos para este tipo de generalización:

- "**Uno frente a uno**": Se entrena un clasificador binario por cada par de emociones. Por tanto, serán necesarios 6*5/2 clasificadores, es decir, 2x6 valores de la regla de clasificación para cada objeto.
- "**Uno frente a todos**": Es necesario entrenar tantos clasificadores binarios como número de emociones se dispongan, en este caso 6. En este caso serán necesarios 6 clasificadores, es decir, 6 valores de la regla de clasificación para cada objeto.

2.7. Aplicaciones

Los sistemas automáticos de reconocimiento de emociones tienen una amplia variedad de aplicaciones. Se puede hacer una diferenciación entre dos campos principales a los cuales se orientan este tipo de aplicaciones:

2.7.1. Aplicaciones que ayudan a mejorar la calidad de vida

- "Call Centers". Son servicios automáticos que detectan emociones y que son capaces de adaptar su respuesta con el locutor o incluso, desviar el control a una persona [28].
- Aplicaciones orientadas a la industria del juego y entretenimiento.
- Sistemas de síntesis de habla emocional para discapacitados.
- Sistemas de automóviles inteligentes capaces de detectar fatiga en el conductor.

El fallo que se detecta en este tipo de aplicaciones es que únicamente disponen de una voz neutra que al locutor puede resultarle aburrida y monótona. El implantar una voz personalizada en estas aplicaciones podría ser de gran utilidad a personas con algún tipo de discapacidad.

2.7.2. Aplicaciones que sirven para mejorar investigaciones relacionadas con la emoción

Algunos de los campos que se dedican al estudio del efecto de las emociones son psiquiatría, psicología o neurología, donde el reconocimiento de emociones puede ayudar en las investigaciones de la conducta social y emocional, en las relaciones familiares, depresiones, trastornos psiquiátricos, etc. [29].

2.7.3. Aplicaciones existentes

El reconocimiento automático de emociones está cogiendo tal importancia que incluso se están desarrollando aplicaciones para tablets y smartphones. Algunos ejemplos de ellas son:

- "**Proyecto emociones**". Se trata de una aplicación que ayuda al desarrollo de la empatía en los niños con autismo. Está desarrollada para dispositivos Android y para Windows. Está desarrollada por un estudiante de último año de la Universidad de Valparaíso, Chile [E9].
- "**MEIT (The Mobile Intelligence Test Emocional)**". Es una aplicación desarrollada por la empresa "Emotional Apps", y permite evaluar las habilidades para percibir, comprender y manejar las emociones [E10].
- "**PicFeel**". Es una aplicación denominada "El Instagram de las emociones", desarrollada por la empresa "Spin Off", cuya idea es tener una red social donde se puedan compartir las emociones y geolocalizarlas mediante fotos [E11].
- "**RFuzzy**". Se trata de una aplicación desarrollada por la facultad de Informática de la Universidad Politécnica de Madrid" que reconoce emociones humanas a través de la voz [E12].
- "**Emotion Sense**". Es una aplicación para Android que se puede descargar desde la aplicación "Google Play". El Smartphone va recogiendo en un segundo plano datos durante el día de una persona, para más tarde sacar resultados sobre su estado de ánimo [E13].
- "**In Flow - Mood and Emotion Diary**". Es una aplicación gratuita que explica diariamente el estado de ánimo de un individuo, dando consejos para ser feliz [E14].

Desarrollar un sistema que nos permita reconocer nuestras propias emociones nos puede servir personalmente para controlar nuestra inteligencia emocional. Es decir, si nosotros mismos somos conscientes de la emoción que sentimos en un momento determinado, podemos llegar a controlarla, intentando que no nos afecte o modificando nuestro comportamiento.

Capítulo 3

Sistema automático de reconocimiento de emociones

En este capítulo se explica detalladamente el sistema automático de reconocimiento de emociones implementado en este proyecto. Además se explican los tipos de características que se han extraído: las características prosódicas y las características espectrales. Las características prosódicas utilizadas en este trabajo se derivan de la frecuencia fundamental, de la intensidad, de la frecuencia del primer formante, y de medidas de calidad de voz. Las características espectrales se componen de los coeficientes mel-cepstrales ("*Mel-Frequency Cepstral Coefficients", MFCC*).

3.1. Introducción

Un patrón se define como una colección de descriptores con los que se pueden representar los principales rasgos de una clase. Un sistema de reconocimiento de patrones es, por tanto, una técnica que mediante el análisis de un conjunto de características de un elemento, asigna una etiqueta que representa a una clase.

Este proyecto se centra en una aplicación particular del reconocimiento de patrones: el reconocimiento automático de emociones. Este tipo de sistemas se basan

en el análisis de características de las emociones para poder clasificar unas frente a otras.

La estructura básica que sigue un sistema automático de reconocimiento de patrones puede observarse en la siguiente figura 15.

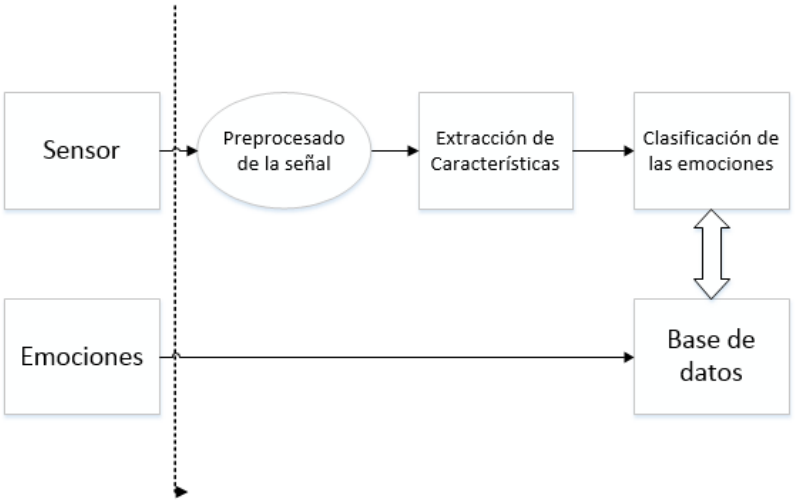

Figura 15. Sistema automático de reconocimiento de patrones

A continuación se describen paso a paso los elementos que forman parte de este sistema.

3.2. Base de Datos

En este trabajo se ha tomado como referencia la base de datos "Berlin Data Base". Como ya se ha explicado en el capítulo 2.5.2, esta base de datos contiene ficheros de audio donde diferentes actores expresan 7 emociones distintas a través de 10 frases.

Las emociones que se representaron fueron: la ira, el miedo, la alegría, la tristeza, el asco, el aburrimiento y la emoción neutra. Para este trabajo la emoción neutra no se ha tenido en cuenta, siendo descartados sus ficheros de audio en cada uno de los experimentos.

3.3. Sensor y preprocesado de la señal

El micrófono contiene un elemento llamado sensor que captura la señal de voz. Una vez que el micrófono recibe la señal acústica, la convierte en señal eléctrica para después procesarla. En primer lugar, se realiza un proceso de conversión analógico-digital, en el que la señal es cuantificada y muestreada.

El siguiente paso consiste en la parametrización de la señal de voz o extracción de sus características acústicas. Se codifica la señal para que el sistema de reconocimiento sea capaz de medirla y evaluarla cuantitativamente. Como etapa previa a la parametrización se puede realizar un preprocesado de la señal, donde se transforma utilizando diversos filtros con el objetivo de facilitar su parametrización o hacerla más eficiente.

En el caso de este trabajo, todos estos pasos previos a la extracción de características fueron realizados por los técnicos que ayudaron en la grabación de la "Berlin Data Base". Algunos de los detalles fueron explicados en el capítulo 2.5.2. En resumen, cada muestra de voz se representa con 16 bits, y la frecuencia de muestreo utilizada es 16 kHz.

3.4. Extracción de características

En este proyecto, se van a extraer tanto características espectrales como prosódicas. Ambos tipos de características se extraen trama a trama y posteriormente son combinadas a un nivel superior. En concreto, se van a considerar dos tipos de niveles: expresión (frase) y clase acústica. En resumen, se diferencian cuatro conjuntos de parámetros acústicos, variando el tipo de características (espectral o prosódica) y la región de expresión sobre la que se calculan (nivel de expresión o nivel de clase), tal y como se muestra en la tabla 7.

La parametrización o extracción de características es idéntica tanto para los datos de entrenamiento como para los datos de verificación.

Granularidad	Tipo de Característica	
	Características prosódicas a nivel de clase	Características espectrales a nivel de clase
	Características prosódicas a nivel de expresión	Características espectrales a nivel de expresión

Tabla 7. Tipos de características

Las características a nivel de expresión o frase corresponden con los estadísticos (media, desviación típica, etc.) de los parámetros espectrales o prosódicos a nivel de trama, calculados sobre toda la frase.

En el caso de las características a nivel de clase, los estadísticos se calculan sobre los segmentos de voz correspondientes a las clases acústicas consideradas. En nuestro caso, los fonemas se agrupan en tres tipos de clases de interés: consonantes, vocales átonas y vocales acentuadas. Los parámetros finales que representan a toda la expresión corresponden con la concatenación de los estadísticos de las tres clases consideradas. La partición de fonemas en distintas clases acústicas más amplias reduce la dependencia de las características con el contenido específico del enunciado, proporcionando robustez y evitando los problemas derivados de la escasez de algunos fonemas en la frase a reconocer.

Para el cálculo de las características a nivel de clase, es necesario realizar una segmentación previa de la frase a nivel de fonema. En la base de datos “Berlín Data Base” se proporciona el alineamiento manual de la transcripción fonética con la voz de cada uno de los ficheros, por lo que este paso inicial no es necesario. Sin embargo, si se quisiera realizar este proceso de forma automática, una opción posible sería realizar una segmentación automática de las frases en fonemas utilizando, por ejemplo, modelos ocultos de Markov.

A continuación, se describen las características a nivel de trama a partir de las cuales se obtienen los parámetros a nivel de expresión o clase.

3.4.1. Características espectrales

Las características espectrales que se extraen en este trabajo son los coeficientes cepstrales en escala de Mel (MFCC) [E15].

$$Cepstrum(s[n]) = \hat{s}[n] = F^{-1}\left[\log\left(\left|F\left[s[n]\right]\right|\right)\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log(\left|S(e^{j\omega})\right|)d\omega$$

Ecuación 17. Coeficientes Cepstrum

Donde s[n] es la convolución entre la excitación y el tracto vocal:

$$s[n] = e[n] * h[n]$$

Ecuación 18. Convolución entre excitación y tracto vocal

Que es lo mismo que:

$$\hat{s}[n] = \hat{e}[n] + \hat{h}[n]$$

Ecuación 19. Convolución en el dominio cepstral

El concepto del Cepstrum se utiliza en muchas técnicas de extracción de características de la señal de voz, aunque es muy vulnerable a los efectos del canal y del ruido aditivo.

Algunos estudios han demostrado que el sistema auditivo humano procesa la señal de voz en el dominio espectral, que se caracteriza por ser más sensible a las bajas frecuencias. Con la escala Mel lo que se consigue es dar más relevancia a las bajas frecuencias, asemejándose de esta manera al sistema auditivo humano y más concretamente, al oído interno. Representando el comportamiento frecuencial del oído humano se puede obtener una mayor eficacia en el sistema de reconocimiento.

En la figura 16 se puede observar el diagrama básico para la extracción de los MFCC.

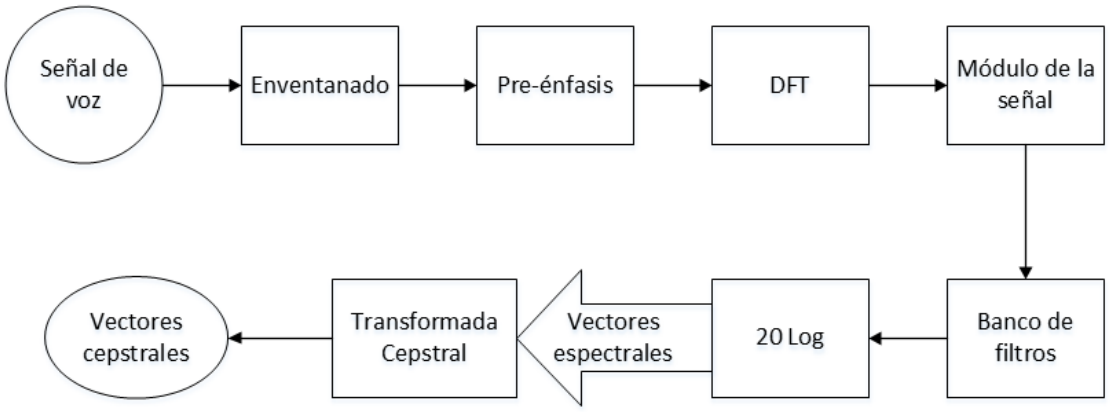

Figura 16. Extracción de los coeficientes MFCC

A continuación se van a analizar con detalle cada uno de los bloques que se representan en la figura anterior.

Enventanado

Que la señal de voz sea un proceso no estacionario supone un problema a la hora de analizarla, pero se puede tener en cuenta la señal a corto plazo (del orden de ms) donde prácticamente la señal se comporta como estacionaria. Para conseguir esto hay que realizar un análisis localizado donde se obtengan tramas o segmentos de la señal de algunos "ms" de duración.

Se denomina enventanado al proceso de generar tramas o segmentos consecutivos de la señal de voz. Por lo general se trabaja con ventanas de tipo Hamming y de tipo Hanning de un tamaño de 20 ms. Con el objetivo de no perder información de la señal de voz, el enventanado se realiza con segmentos solapados entre sí. El solapamiento entre las ventanas suele ser de 10 ms, obteniéndose por tanto, coeficientes MFCC cada 10 ms.

Pre-énfasis

El objetivo de este filtro por el que pasa la señal de voz, es compensar la atenuación de aproximadamente -20db/década que el mecanismo de producción del habla genera. Este paso no es obligatorio, pero si se recomienda si se quieren enfatizar los formantes de alta frecuencia.

El filtro de pre-énfasis es de la siguiente manera:

$$y[n] = x[n] - \alpha x[n-1]$$

Ecuación 20. Filtro pre-énfasis

Donde α toma valores del intervalo [0.95, 0.98].

Transformada discreta de Fourier (DFT)

Una vez se ha realizado el enventanado de la señal, se calcula la DFT de ficha señal:

$$x[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk} \qquad 0 \leq k \leq N$$

Ecuación 21. Transformada discreta de Fourier

Es en este momento donde se deja a un lado la fase y sólo se trabaja con el módulo de la señal.

Banco de filtros

El módulo de la señal x[k] se multiplica por un banco de filtros triangulares de área unidad, como el de la figura 17, los cuales están espaciados de acuerdo con las frecuencias de la escala Mel.

$$mel(f) = 2595\, log_{10}\left(1 + \frac{f}{700}\right)$$

Ecuación 22. Escala de frecuencias Mel

Donde f es la frecuencia representada en el eje lineal.

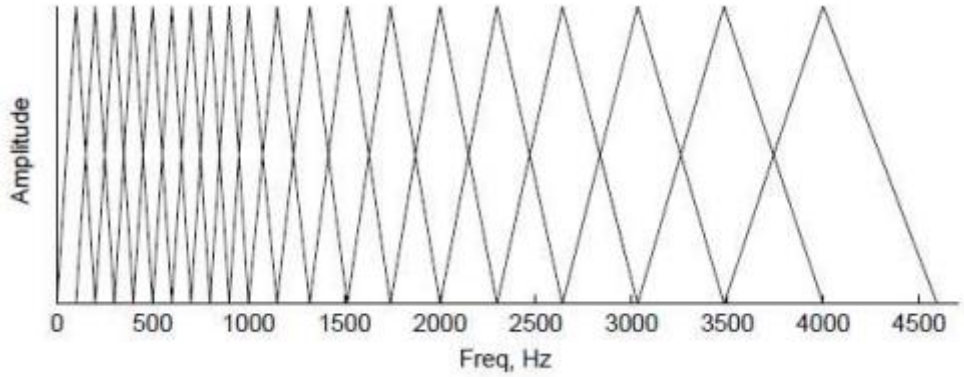

Figura 17. Banco de filtros triangulares de área unidad, [E15]

El ancho de banda de estos filtros triangulares viene determinado por la frecuencia central de cada filtro, la cual depende del número de filtros y de la frecuencia de muestreo. El ancho de banda de cada uno de los filtros aumentará conforme disminuya el número de filtros del banco.

En la figura 18 se representa la escala Mel frente a la lineal para poder observar la relación entre ambas.

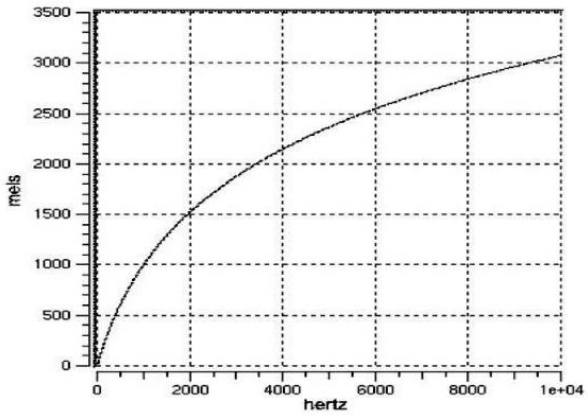

Figura 18. Escala de Mel, [E15]

Es de gran utilidad tener una expresión matemática que permita calcular el valor de cada uno de los filtros:

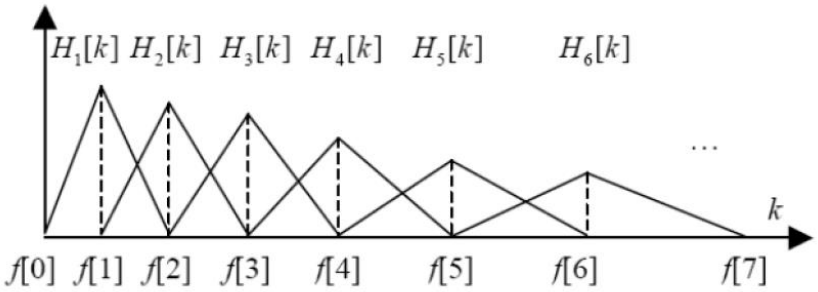

Figura 19. Banco de filtros Mel, [E15]

$$H_m[k] = \begin{cases} 0 & ,k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & ,f[m-1] \le k \le f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & ,f[m] \le k \le f[m+1] \\ 0 & ,k > f[m+1] \end{cases}$$

Ecuación 23. Expresión matemática para calcular el valor de los filtros Mel

Donde $1 \le \mathrm{m} \le \mathrm{F}$, y F representa al número de filtros del banco. Además se tiene que:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}\right)$$

Ecuación 24. Calculo de los extremos de los filtros triangulares

$$B^{-1}(b) = 700\ (e^{\frac{b}{2595}} - 1)$$

Ecuación 25.

Considerando f_h y f_1 los extremos superior e inferior de cada filtro triangular.

Cálculo de las energías en banda

Se calcula la energía de cada uno de los filtros después de haber multiplicado el banco de filtros por el módulo de la transformada de Fourier de la señal de voz. La forma de calcular la energía se puede apreciar en la ecuación 26.

$$E_m = \sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \qquad 1 \leq m \leq F$$

Ecuación 26. Cálculo de la energía

Una vez que se ha calculado la energía de cada filtro, se calcula su logaritmo. Uno de los inconvenientes de trabajar con energías en banda es que los espectros de las bandas adyacentes de cada uno de los filtros tienen una alta correlación, obteniéndose coeficientes espectrales que son dependientes estadísticamente entre ellos.

Transformada discreta del coseno ("Discrete Cosine Transform", DCT)

La técnica que se utiliza para eliminar esa correlación estadística es calcular la transformada discreta del coseno, donde se transforman los coeficientes espectrales al dominio cepstral, dando lugar a lo que se conoce como coeficientes cepstrales o MFCC.

$$C_{MFCC}[m] = \sum_{k=0}^{N-1} \log(E_k) \cos\left(m\left(k - \frac{1}{2}\right)\frac{\pi}{N}\right) \qquad m = 1, \ldots, F$$

Ecuación 27. Coeficientes MFCC

Además de los MFCC, se pueden calcular los coeficientes Delta-MFCC (ΔMFCC) y Delta-Delta-MFCC (ΔΔMFCC) que corresponden respectivamente con la primera y segunda derivada de los MFCC y aportan información sobre la coarticulación de los fonemas. Dichos coeficientes ofrecen robustez al sistema de reconocimiento frente a la variabilidad del interlocutor.

Los Delta-MFCC son conocidos como coeficientes de velocidad debido a que miden la variación de los MFCC en un instante de tiempo. Por la misma razón, a los

Delta-Delta-MFCC se les llama coeficientes de aceleración, ya que muestran la variación de Delta-MFCC en un instante de tiempo. Estos coeficientes se expresan matemáticamente como muestra la ecuación 28.

$$\begin{cases} \Delta c_{MFCCi}[m] = \sum_{k=-l}^{l} \frac{k c_{MFCC(i+l)}[m]}{|k|} \\ \Delta\Delta c_{MFCCi}[m] = \sum_{k=-l}^{l} \frac{k \Delta c_{MFCC(i+l)}[m]}{|k|} \end{cases} \qquad 1 \leq m \leq F$$

Ecuación 28. Coeficientes Delta-MFCC y Delta-Delta_MFCC

Siendo l el parámetro que controla el número de tramas y por tanto, el intervalo de tiempo en el que se calcula la información.

3.4.2. Características prosódicas

La prosodia puede definirse como la rama de la lingüística que analiza y representa formalmente aquellos elementos de la expresión oral, tales como el acento, los tonos y la entonación.

A continuación se explican las características prosódicas que se han extraído de la señal de voz en este proyecto, para después analizarlas y compararlas.

3.4.2.1. Frecuencia fundamental o "Pitch"

El pitch es la frecuencia fundamental a la que vibran las cuerdas vocales, también llamada frecuencia fundamental o F_0. La frecuencia fundamental es la frecuencia más baja del espectro de frecuencias, de tal forma que las frecuencias dominantes pueden expresarse como múltiplos de esta. Las características de la frecuencia fundamental contienen mucha información emocional y permiten caracterizar la voz del hablante [31].

- El valor medio del Pitch expresa el nivel de excitación del hablante y depende de éste. Si la media del Pitch es un valor elevado, es indicador de que puede haberse producido un grado mayor de excitación.
- El rango del Pitch se define como la distancia entre los valores mínimo y máximo de la frecuencia fundamental. Este parámetro muestra el grado de exaltación del hablante. Un rango más amplio que el normal refleja una excitación psicológica o emocional.

- Las fluctuaciones en el Pitch se describen como la velocidad de las transiciones entre valores altos y bajos. En general, la curva de tono es discontinua para las emociones consideradas como negativas (enfado, miedo), y es suave para las emociones positivas (alegría).
- Representando la distribución del Pitch se puede describir el rango de valores del Pitch así como la probabilidad de que un cierto valor esté dentro de un subconjunto de dicho rango. Esta es una de las formas más fáciles de distinguir entre voz masculina y femenina, ya que las voces femeninas tienen una frecuencia fundamental media mayor a la del hombre, además de una desviación típica mucho mayor, figura 20.

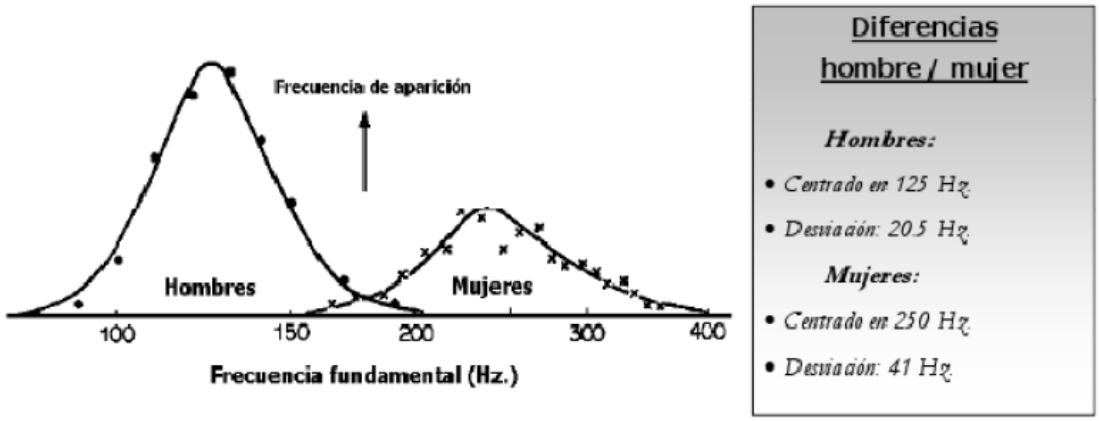

Figura 20. Frecuencia fundamental en hombres y mujeres, [30]

3.4.2.2. Formantes

Los formantes son las frecuencias de un sonido donde se concentra la mayor parte de la energía sonora, es decir, son las frecuencias naturales de vibración del tracto vocal que permiten caracterizar un sonido frente a los demás.

El primer formante, F_1, depende directamente de la apertura de la mandíbula. Cuanto más abierta esté la mandíbula, más alta será la frecuencia de F_1. Además, el valor de este formante varía inversamente proporcional a la altura de la lengua, es decir, cuanta más alta se sitúe la lengua, menor será el valor de la frecuencia del primer formante.

Por otro lado, el segundo formante, F_2, varía con la posición de la lengua. Cuando la lengua se sitúa hacia el interior de la cavidad oral, el valor del segundo formante asciende.

3.4.2.3. Intensidad de la voz

La intensidad de la voz se define como la energía o potencia acústica transmitida por segundo. Este parámetro es capaz de expresar rasgos emocionales.

La intensidad de la voz permite detectar estados de ánimo, incluso describe algunos aspectos del carácter del interlocutor. Características como la agresividad, el miedo, el nerviosismo o la tensión se detectan con un volumen alto de intensidad; mientras que la depresión, la tristeza o el cansancio suelen asociarse con un volumen más bajo de intensidad [32].

La intensidad se mide en decibelios. El nivel medio de una conversación ronda en torno a los 50 dB.

Los parámetros jitter, shimmer y HF500 que se explican a continuación son medidas prosódicas de la calidad de la voz [33].

3.4.2.4. Jitter

El jitter se define como las variaciones de la frecuencia fundamental que aparecen en los tramos sonoros de la señal de voz, es decir, es la fluctuación periodo a periodo de la frecuencia fundamental. El jitter describe un ruido por modulación en frecuencia.

En la figura 21 se puede observar la variabilidad de F0.

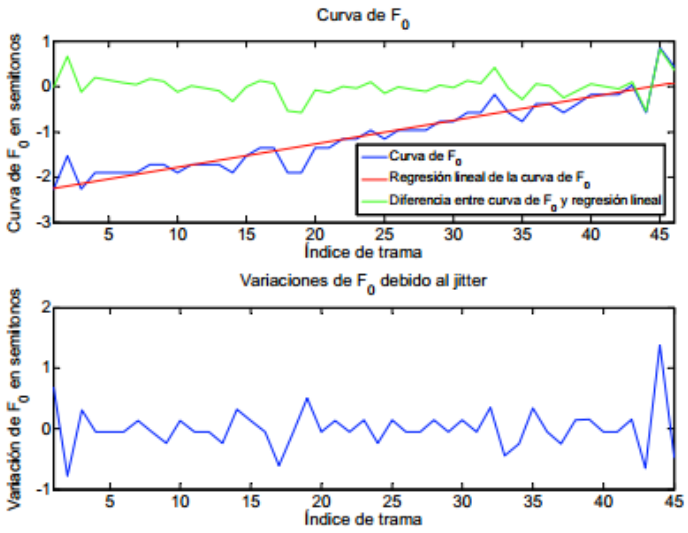

Figura 21. Jitter, [33]

3.4.2.5. Shimmer

El shimmer, a diferencia de jitter, describe un ruido por modulación en amplitud. En la figura 22 se observa como este parámetro calcula las variaciones de amplitud de la forma de onda en los tramos sonoros de la señal de voz.

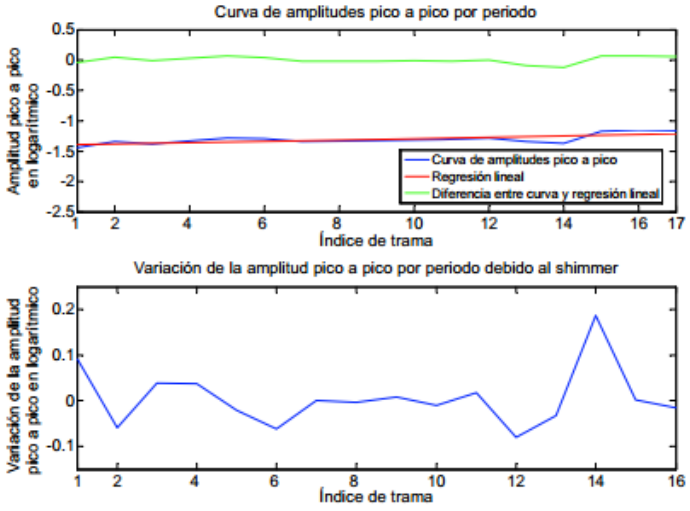

Figura 22. Shimmer, [33]

En la figura 23 se representa un análisis de los parámetros jitter y shimmer de 5 emociones diferentes.

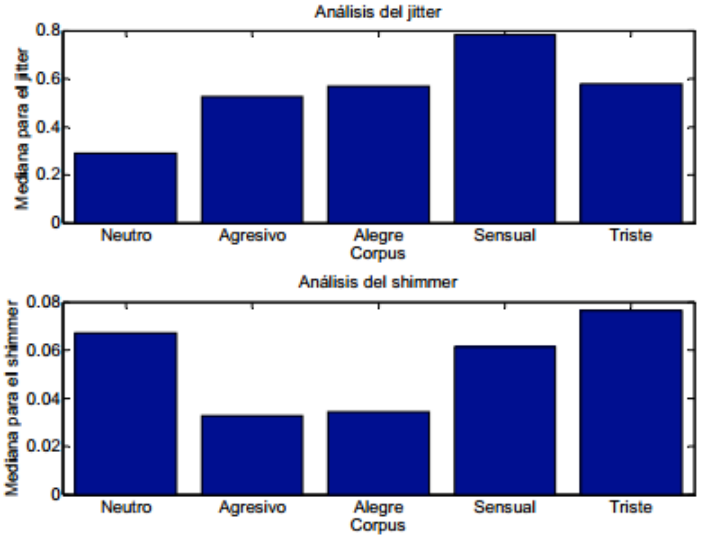

Figura 23. Comparación de Jitter y Shimmer en emociones, [33]

3.4.2.6. HF500

El parámetro HF500 es una medida de la calidad de voz que representa la relación de la energía en alta frecuencia con la energía en baja frecuencia, alrededor de una frecuencia de corte de 500 Hz. La incorporación del parámetro HF500 produce un aumento en la tasa de identificación de la emoción “enfado” y disminuye la tasa de la emoción “triste”.

3.4.2.7. Duración

La velocidad del habla y la situación de los acentos describen una componente de la prosodia llamada duración, cuyos efectos son el ritmo y la velocidad. El ritmo en el habla viene determinado por la situación de los acentos y por la combinación de las duraciones de las pausas y de los fonemas.

En algunas condiciones de estrés, los cambios entre vocales y consonantes, la presencia de consonantes o la duración de las palabras tienen mucha importancia a la hora de que el oyente pueda interpretar de manera correcta la información emocional del locutor [34].

Existen una serie de parámetros que dependen de la duración y que ayudan a distinguir algunas emociones:

- **Velocidad de locución**: Por lo general, un locutor excitado acorta la duración de las sílabas, de esta manera la velocidad de locución medida en sílabas por segundo o en palabras por minuto se incrementa.
- **Número de pausas y su duración**: Un locutor en un estado de exaltación tiende a hablar de una forma rápida, haciendo menos pausas y más cortas, mientras que un locutor deprimido o triste tiende a hablar más lentamente, realizando unas pausas más largas.
- **Cociente entre el tiempo de locución y el de las pausas**.

3.5. Clasificación de emociones

Un sistema automático de reconocimiento de emociones tiene el objetivo de asignar una de las emociones anteriormente predefinidas a cada una de las muestras de voz que han sido proporcionadas, minimizando el error promedio de clasificación de cara a futuras observaciones.

Se pueden definir dos etapas en el diseño de un clasificador: la fase de entrenamiento y la fase de reconocimiento o test.

En la fase de entrenamiento se utilizan las que se han considerado muestras de entrenamiento para construir el clasificador, considerando las restricciones que pudiera haber. Una vez que se dispone de la regla de clasificación, por la que se asigna cada fichero de voz a una emoción, se pasa a la fase de reconocimiento. En esta fase se utilizan las muestras que no han sido utilizadas en la fase de entrenamiento.

Para los experimentos que se van a realizar en este trabajo, se va a utilizar el clasificador SVM explicado en el capítulo 2, implementado con la librería LIBSVM. LIBSVM se trata de una librería para máquinas de vectores soporte ("Support Vector Machine, SVM"). SVM es un método de aprendizaje automático para la clasificación, regresión y otras tareas de aprendizaje. Su software se puede encontrar en [E16].

El proceso que utiliza SVM se divide en las dos fases que se han explicado anteriormente. Primero se entrena un conjunto de datos para obtener un modelo y después se usa ese modelo que se ha construido para predecir información sobre un conjunto de datos de reconocimiento, distinto al de entrenamiento [35].

3.5.1. BAC (Balanced ACcuracy)

Para medir las prestaciones del sistema, y dado que el número de expresiones o frases grabadas es distinto para cada emoción, se usa el BAC ("Balanced ACcuracy") como indicador de rendimiento para el reconocimiento de emociones en los experimentos realizados. BAC se define como el promedio de tasa de acierto sobre todas las clases de emociones, ponderado por el número de frases que hay de cada emoción.

$$BAC = \frac{1}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}$$

Ecuación 29. BAC

Donde K es el número de clases de emociones, N_i es el número total de expresiones pertenecientes a la emoción i y n_i es el número de expresiones pertenecientes a la clase que se está evaluando.

Cuando se usa una clasificación estándar, se tiene en cuenta el número de expresiones clasificadas correctamente con respecto al total, sin embargo la técnica BAC no es sensible al desequilibro de la distribución de expresiones entre las clases. Un ejemplo de esto es considerar una clasificación binaria entre las emociones “alegría” y “miedo” en un conjunto de datos donde 90 ficheros corresponden a la emoción “alegría” y los 10 restantes a la emoción “miedo”. Si se hace una predicción de las emociones con respecto a la clase mayoritaria, el resultado sería un 90% para la clase “alegría”, mientras que si se utiliza la técnica BAC su resultado sería el 50%.

3.5.2. Leave-One-Subject-Out (LOSO)

Con el fin de obtener estabilidad e independencia de los hablantes en los clasificadores que se van a obtener, la fase de reconocimiento se va a realizar mediante la técnica validación cruzada dejando uno fuera (“Leave-One-Subject-Out”, LOSO). Esta técnica separa los datos de manera que en cada iteración únicamente se tiene un dato de prueba, en éste caso, los ficheros pertenecientes a una sola clase, y el resto de ficheros se usan para entrenamiento. Gracias a esto se puede asegurar que el conjunto de ficheros usados para la fase de reconocimiento no fueron utilizados para la fase de entrenamiento.

Capítulo 4

Pruebas experimentales y resultados

En este capítulo se explican los experimentos propuestos en el capítulo anterior con las diferentes pruebas que se han realizado para cada uno de ellos, para después hacer una pequeña comparación de los principales resultados obtenidos.

4.1. Introducción

La extracción de características a partir de ficheros de voz es la parte principal de este trabajo. En el capítulo 3 se han descrito los tipos de características que se van a extraer:

- Características espectrales a nivel de expresión
- Características espectrales a nivel de clase
- Características prosódicas a nivel de expresión
- Características prosódicas a nivel de clase

En los diferentes experimentos desarrollados a continuación se explica cómo se han extraído cada una de estas características y el conjunto de pruebas realizadas para llegar al mejor resultado.

Para cada experimento se han realizado diferentes pruebas modificando algunos de los parámetros utilizados para la extracción de cada tipo de características.

Cada prueba consiste en un conjunto de 10 subexperimentos, que a su vez contienen una lista con ficheros de entrenamiento y otra con ficheros de reconocimiento previamente estructuradas. En concreto, en cada subexperimento el sistema se entrena con nueve de los diez locutores disponibles en la base de datos y se evalúa con el locutor restante. Los resultados presentados corresponden con la media de los obtenidos en los 10 subexperimentos antes mencionados.

El diagrama de flujo del sistema implemento en Matlab se corresponde con el de la figura 24.

4.2. Extracción de características espectrales a nivel de expresión

Las características espectrales a nivel de expresión son los estadísticos (valores medios, las desviaciones estándar, etc.) de los MFCC calculados sobre toda la frase. Para cada expresión, se calculan 12 MFCC con una ventana de Hamming de 25 ms a intervalos de 10 ms. En algunos experimentos se añade el coeficiente cesptral de orden 0 y en otros, se sustituye por la log-energía de cada trama, según se indica en la tabla 9. Además, se calculan los coeficientes velocidad y aceleración como la primera y segunda derivada de los MFCC utilizando las diferencias finitas, como ya se explicó en el capítulo 3. Esto da lugar a vectores de características a nivel de trama formados por 39 parámetros.

Finalmente, se calcula la media, la desviación estándar y, en algunos experimentos, la asimetría (“skewness”) de dichos vectores, de modo que el número total de características espectrales a nivel de expresión es de 78 en caso de utilizar la media y la desviación estándar, y de 117 en caso de añadir la asimetría a los estadísticos anteriores. Estas características son la entrada al clasificador, que en nuestro caso está basado en SVMs.

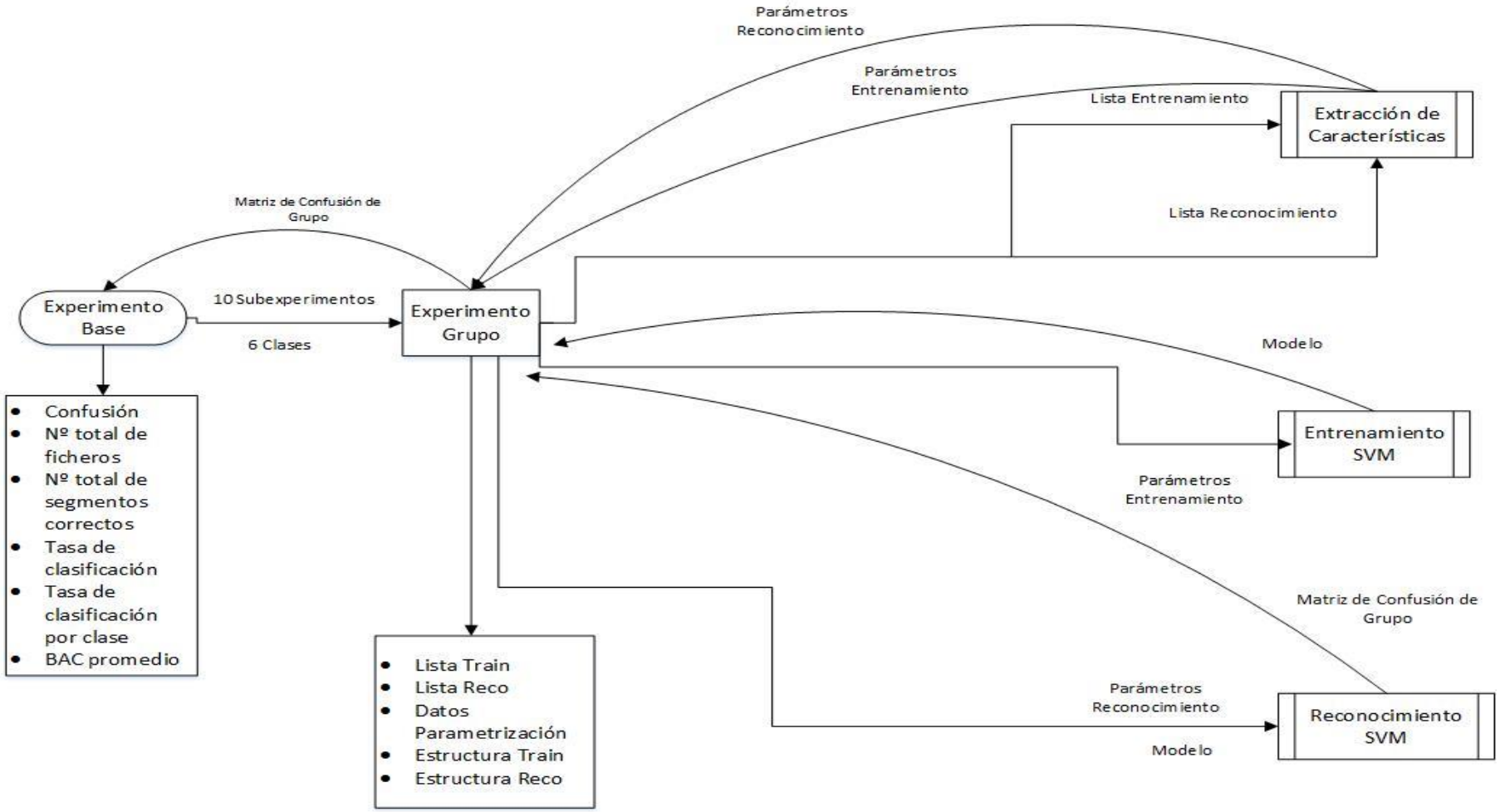

Figura 24. Diagrama de flujo del sistema de reconocimiento de emociones

Para extraer los MFCC se utiliza la función “melcepst” perteneciente al paquete “voicebox” implementado en Matlab. Esta función calcula los coeficientes “Mel Cepstrum” de una señal de voz determinada y tiene la forma:

```
[c] = melcepst(x, fs, w, nc, p, n, inc);
```

El parámetro “x” es la señal de voz sobre la que se van a calcular los coeficientes, “fs” es la frecuencia de muestreo de los ficheros de voz y los términos de “w” pueden variar dependiendo del tipo de parámetros que se quieran extraer.

Las diferentes combinaciones de “w” que se han realizado vienen determinadas por los siguientes parámetros:

- ‘0’: Incluye el coeficiente de orden 0 de los coeficientes espectrales.
- ‘E’: Incluye el coeficiente log-energía.
- ‘d’: Incluye el coeficiente delta (primera derivada, dc/dt).
- ‘D’: Incluye el coeficiente delta-delta (segunda derivada, d^2c/dt^2).

El resto de valores que se han utilizado en la parametrización se muestran en la tabla 8:

Característica	Sigla	Valor
Frecuencia de muestreo	fs	16000
Número de coeficientes	nc	12
Número de filtros	p	40
Tamaño de trama (ms)	nt	25
Solape (ms)	inct	10

Tabla 8. Parámetros parametrización

Para ello es necesario hacer algunas conversiones de los tamaños de ventana y solapes a tramas:

- n = round(fs*nt*1e-3)
- inc = round(fs*inct*1e-3)

Una vez calculados los MFCC a nivel de trama con sus diferentes parámetros, se calculan sobre toda la frase diversas medidas estadísticas como son la media, la desviación típica y el “skewness” (estadístico de orden 3) de dichos coeficientes.

En la tabla 9 se pueden observar las distintas pruebas realizadas con la combinación de parámetros calculados para cada una de ellas.

EXPERIMENTO 1	Prueba 1	Prueba 2	Prueba 3	Prueba 4	Prueba 5	Prueba 6	Prueba 7	Prueba 8
E --> Log-Energy	Si	Si	No	No	Si	Si	No	No
0 --> Coeficiente de orden 0	No	No	Si	Si	No	No	Si	Si
d --> Coeficiente delta	Si	Si	Si	Si	Si	Si	Si	Si
D --> Coeficiente delta-delta	Si	No	Si	No	Si	No	Si	No
Media	Si	Si	Si	Si	Si	Si	Si	Si
Desviación Típica	Si	Si	Si	Si	Si	Si	Si	Si
Skewness	No	No	No	No	Si	Si	Si	Si
BAC promedio (%)	62,5	61,67	63,56	64,44	60,65	58,85	64,03	62,82
Tasa de clasificación promedio (%)	65,64	64,76	65,86	66,74	65,2	63,44	68,06	66,96

Tabla 9. Pruebas del experimento 1

Aunque la tasa de clasificación promedio es mayor en la prueba 7 con un 68,06% de acierto, en la prueba 4 se ha obtenido el mejor BAC promedio correspondiente a un 64,44% de acierto, siendo éste el valor que se utilizará para comparar el resultado de los distintos sistemas, ya que como se explicó en el capítulo 3, el BAC tiene en cuenta que en el test hay distinto número de ficheros por emoción.

En la prueba 4 se han extraído los MFCC incluyendo el coeficiente cepstral de orden 0 y los coeficiente delta, es decir, la primera derivada de los MFCC. Los parámetros que se han utilizado para la clasificación se han formado con la media y la desviación típica de los coeficientes obtenidos.

A continuación, en la tabla 10 se muestra la matriz de confusión correspondiente a la prueba 4.

		Emoción identificada					
		Felicidad	Tristeza	Ira	Miedo	Asco	Aburrimiento
Emoción interpretada	Felicidad	55	2	5	-	4	4
	Tristeza	6	32	-	21	9	14
	Ira	6	-	51	-	5	1
	Miedo	1	20	0	101	9	2
	Asco	4	7	5	5	40	1
	Aburrimiento	7	10	1	-	2	24
	Total	79	71	62	127	69	46

Tabla 10. Matriz de confusión de la mejor prueba del experimento 1

Las dos emociones que se reconocen peor son tristeza y aburrimiento. Observando la matriz de confusión mostrada en la tabla [] se puede llegar a la conclusión que la emoción tristeza con 32 ficheros clasificados correctamente, se confunde en 20 ocasiones con la emoción miedo y 10 con la emoción aburrimiento.

Además, la emoción aburrimiento se clasifica correctamente en 24 ocasiones, pero en 14 ocasiones se confunde con la emoción tristeza.

En la figura 25 se muestran las tasas de clasificación promedio de cada una de las clases para este experimento.

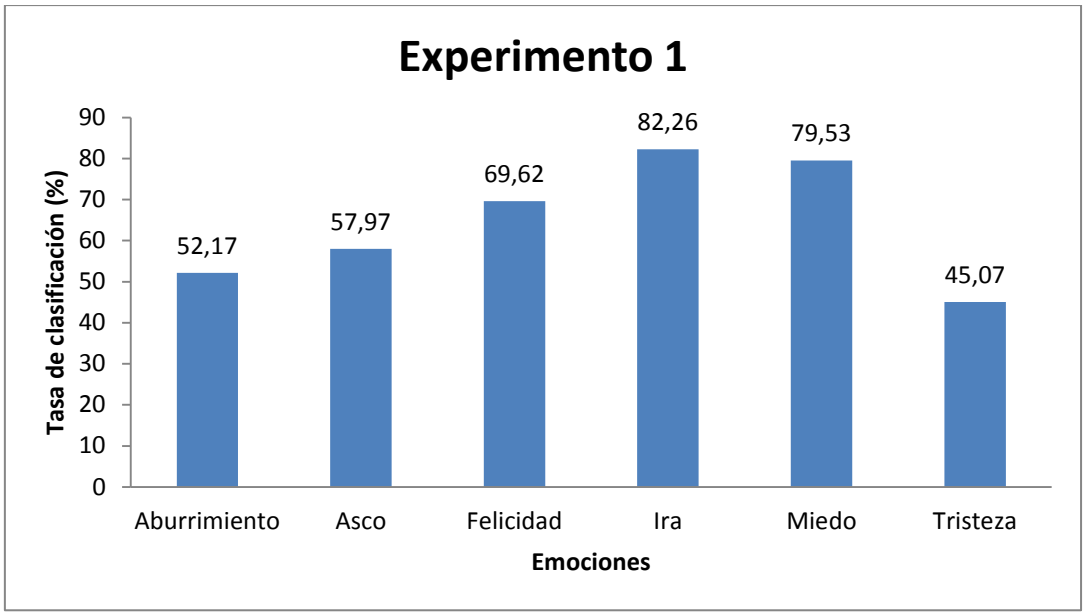

Figura 25. Tasas de clasificación promedio de las emociones del experimento 1

En este primer experimento basado en la extracción de características espectrales a nivel de expresión, se puede sacar en conclusión que la emoción tristeza es la que peor se reconoce con una tasa de clasificación promedio del 45.07%, mientras que la emoción mejor reconocida es la ira con una tasa del 82.26%.

4.3. Extracción de características espectrales a nivel de clase

Estas características espectrales modelan cómo se codifica la emoción en el habla a nivel de fonema. Usando la segmentación de la expresión a nivel de fonema, se forma el vector de características espectrales mediante la concatenación de las medias y las desviaciones estándar de los MFCC de las vocales acentuadas, vocales átonas y consonantes. Además se calcula la media de la duración de cada clase de fonema.

En resumen, el vector de características espectrales a nivel de clase es de 237 dimensiones y consta de los siguientes grupos de características:

- Media y desviación estándar de los MFCC de las vocales acentuadas.
- Media y desviación estándar de los MFCC de las vocales átonas.

- Media y desviación estándar de los MFCC de las consonantes.
- Media de la duración de las vocales acentuadas, átonas y consonantes.

La información sobre la segmentación de las frases a nivel de fonema se obutvo de la segmentación manual disponible en la base de datos “Berlín emotional speech database”, que se encuentra en el directorio de la distribución llamado “lablaut”. Dicho directorio contiene un fichero “.lablaut” con la transcripción fonética de cada fichero de audio “.wav” perteneciente a dicha base de datos.

Cada fichero “.lablaut” tiene el mismo formato, el cual se puede apreciar en la figura 26:

```
signal wochenAO
type 0
comment created using xlabel Wed Apr 26 12:59:18 2000
font -misc-*-bold-*-*-*-15-*-*-*-*-*-*-*
separator ;
nfields 1
#
     0.080109   -1 _t
     0.119160   -1 !t   +asp
     0.144370   -1 E6
     0.244223   -1 l
     0.348031   -1 a          2
     0.427091   -1 p
     0.509149   -1 !p   +na
     0.551660   -1 m    +sil
     0.602575   -1 l
     0.698978   -1 i
     0.768183   -1      +hoe
     0.802292   -1 t
     0.846286   -1 !t
     0.865565   -1 a    +lar
     0.903133   -1      -lar
     0.938230   -1 f
     0.981748   -1 d    -sth
     1.025743   -1 !d
     1.044033   -1 I
     1.081601   -1 m
     1.146357   -1 aI   +lar  3
     1.192329   -1      -lar
     1.331737   -1 s    +sth
     1.343601   -1      -sth
     1.392539   -1 S
     1.491403   -1 X
     1.546280   -1 a    +nas  1
     1.645639   -1 N
     1.724730   -1 k    +nas
     1.746975   -1 !k
     1.765265   -1 kasp
     1.841390   -1 .
```

Figura 26. Formato fichero ".lablaut" de "Berlin Data Base"

El texto que aparece al principio es común para todos los ficheros, indica la fecha en la que se creó dicho documento. La primera columna indica el instante inicial de cada símbolo. La segunda columna muestra un código de control, en este trabajo se

obvia este dato. La tercer columna muestra los símbolos fonéticos, determinados por fonetistas, y que más adelante se clasificarán en vocales, vocales acentuadas y consonantes. La cuarta columna indica alguna característica particular de la pronunciación de ciertos símbolos, que puede ser nasalizado, aspirado, etc. La quinta columna diferencia las vocales acentuadas.

Para realizar el experimento, se requiere tener un formato más útil, para que a la hora de leer cada uno de los ficheros ".lablaut" se pueda obtener la información necesaria. Para ello se creó una función en Matlab que modifica cada uno de los ficheros y los deja con el formato de la figura 27:

```
03a01Faxx   0.080109   0.119160   _t
03a01Faxx   0.119160   0.144370   !t
03a01Faxx   0.144370   0.244223   E6
03a01Faxx   0.244223   0.348031   l
03a01Faxx   0.348031   0.427091   a     2
03a01Faxx   0.427091   0.509149   p
03a01Faxx   0.509149   0.551660   !p
03a01Faxx   0.551660   0.602575   m
03a01Faxx   0.602575   0.698978   l
03a01Faxx   0.698978   0.768183   i
03a01Faxx   0.768183   0.802292   i
03a01Faxx   0.802292   0.846286   t
03a01Faxx   0.846286   0.865565   !t
03a01Faxx   0.865565   0.903133   a
03a01Faxx   0.903133   0.938230   a
03a01Faxx   0.938230   0.981748   f
03a01Faxx   0.981748   1.025743   d
03a01Faxx   1.025743   1.044033   !d
03a01Faxx   1.044033   1.081601   I
03a01Faxx   1.081601   1.146357   m
03a01Faxx   1.146357   1.192329   aI          3
03a01Faxx   1.192329   1.331737   aI
03a01Faxx   1.331737   1.343601   s
03a01Faxx   1.343601   1.392539   s
03a01Faxx   1.392539   1.491403   S
03a01Faxx   1.491403   1.546280   X
03a01Faxx   1.546280   1.645639   a           1
03a01Faxx   1.645639   1.724730   N
03a01Faxx   1.724730   1.746975   k
03a01Faxx   1.746975   1.765265   !k
03a01Faxx   1.765265   1.841390   kasp
```

Figura 27. Formato adaptado del fichero ".lablaut"

Para cada una de las filas, se tiene en la primera columna el nombre del fichero con el que se trabaja en ese momento; en la segunda y tercera columna los tiempos de inicio y fin de los símbolos fonéticos; en la cuarta columna se pueden encontrar cada uno de los símbolos de la expresión, y en la quinta columna se muestran los números que diferencian si una vocal está acentuada.

Para hacer la clasificación de fonemas, se recogen todos los símbolos en un solo documento y se determinan cuales son vocales y cuales son consonantes, como se muestra en la tabla 11.

Consonantes	Vocales
!	6
!b	8
!b+na	9
!d	@
!g	@6
!k	@E
!p	@O
!t	@a
4	AO
?	AU
B	E
C	E6
D	I
G	I+:
H	I+nas
J	I6
L	I@
N	O
NN	OU
R	OY
S	U
X	V
_b	Y
_d	Y6
_g	a
_k	a6
_p	aI
_t	aO
b	aU

Consonantes	Vocales
basp	ao
d	e
dasp	e@
dsap	hoe
f	i
g	i6
g-sth	i@
gasp	ia
h	o
j	u
k	y
kasp	
l	
m	
mm	
n	
nn	
p	
pasp	
s	
ss	
t	
tasp	
tasp+sth	
v	
vv	
x	
z	

Tabla 11. Conjunto de símbolos clasificados en vocales y consonantes

Para calcular las características espectrales a nivel de clase se creó una función en Matlab que calcula las medias y desviaciones estándar de los MFCC y las medias de las duraciones del conjunto de vocales átonas, vocales acentuadas y consonantes que aparecen en cada fichero de voz.

Para calcular los MFCC de cada clase se plantearon inicialmente dos alternativas:

1. Convertir los tiempos de inicio y fin de cada símbolo a muestras multiplicándolos por la frecuencia de muestreo, fs = 16000 Hz. Se coge la señal de voz almacenada entre esas muestras y se calculan los MFCC de cada segmento. El inconveniente de este método es que hay que parametrizar la señal de voz por cada símbolo del fichero.
2. Se parametriza el fichero de voz entero y se determinan las tramas que interesan. Después con las marcas de inicio y fin de cada fonema, se cogen los MFCC correspondientes al símbolo fonético en cuestión. Finalmente se eligió este método debido a que es más rápido ya que sólo es necesario parametrizar la señal de voz una vez.

Igual que en el apartado anterior, los coeficientes MFCC de cada fichero se calculan a través de la función melcepst e igualmente se utilizan los mismos datos.

En la tabla 12 se pueden observar las distintas combinaciones de parámetros que se han utilizado en la realización de las pruebas.

EXPERIMENTO 2	Prueba 1	Prueba 2	Prueba 3	Prueba 4	Prueba 5	Prueba 6	Prueba 7	Prueba 8	Prueba 9	Prueba 10	Prueba 11
E --> Log-Energy	Si	Si	No	No	Si	Si	No	No	Si	Si	Si
0 --> Coeficiente de orden 0	No	No	Si	Si	No	No	Si	Si	No	No	No
d --> Coeficiente delta	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
D --> Coeficiente delta-delta	No	Si	Si	No	No	Si	Si	No	No	No	Si
Media	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
Desviación Típica	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
Skewness	No	No	No	No	Si	Si	Si	Si	No	No	No
Duraciones	Si	SI	Si	SI	Si	SI	Si	SI	No	No	No
Juntar Vocales en 1 grupo	No	No	No	No	No	No	No	No	No	SI	Si
BAC promedio (%)	65,41	65,51	64,87	65,1	66,98	67,08	64,78	65,79	65,41	65,06	65,23
Tasa de clasificación promedio (%)	67,7	67,92	67,92	68,14	69,61	70,13	68,58	69,25	67,7	67,26	67,48

Tabla 12. Pruebas del experimento 2

En este caso, la prueba que ha obtenido mejores resultados corresponde a la número 6, donde se han extraído los MFCC con sus coeficientes log-energía, delta y delta-delta. Después se ha calculado la media, desviación típica y skewness de dichos coeficientes a nivel de clase, obteniéndose un BAC del 67,08% y coincidiendo con que la mejor tasa de clasificación promedio también pertenece a esta prueba, con un valor del 70,13%.

En la prueba 9 se han incluido los mismos parámetros que en la prueba 1, exceptuando las duraciones de cada una de las clases, obteniéndose el mismo resultado en ambas pruebas, por lo que se llega a la conclusión de que estas no influyen en la extracción de este tipo de características. Además, la prueba número 10 se ha realizado juntando los dos tipos de vocales en una sola clase y manteniendo el resto de parámetros como los de la prueba 9 y no se ha obtenido un mejor resultado.

La matriz de confusión obtenida en la prueba 6 se muestra en la tabla 13.

Emoción interpretada	Emoción identificada					
	Felicidad	Tristeza	Ira	Miedo	Asco	Aburrimiento
Felicidad	56	-	5	-	9	2
Tristeza	6	37	-	22	9	10
Ira	6	0	47	-	3	1
Miedo	-	19	-	102	6	1
Asco	7	8	10	-	36	5
Aburrimiento	4	7	-	2	5	27
Total	79	71	62	126	68	46

Tabla 13. Matriz de confusión de la mejor prueba del experimento 2

Si se observa la matriz de confusión que se muestra en la tabla 13 se concluye que la emoción tristeza con 37 ficheros clasificados correctamente, se confunde en 19 ocasiones con la emoción miedo. También se puede observar que la emoción aburrimiento se clasifica de una forma correcta en 27 ocasiones, pero se confunde 10 veces con la emoción tristeza.

En la figura 28 se muestran las tasas de clasificación promedio de cada una de las clases de este experimento.

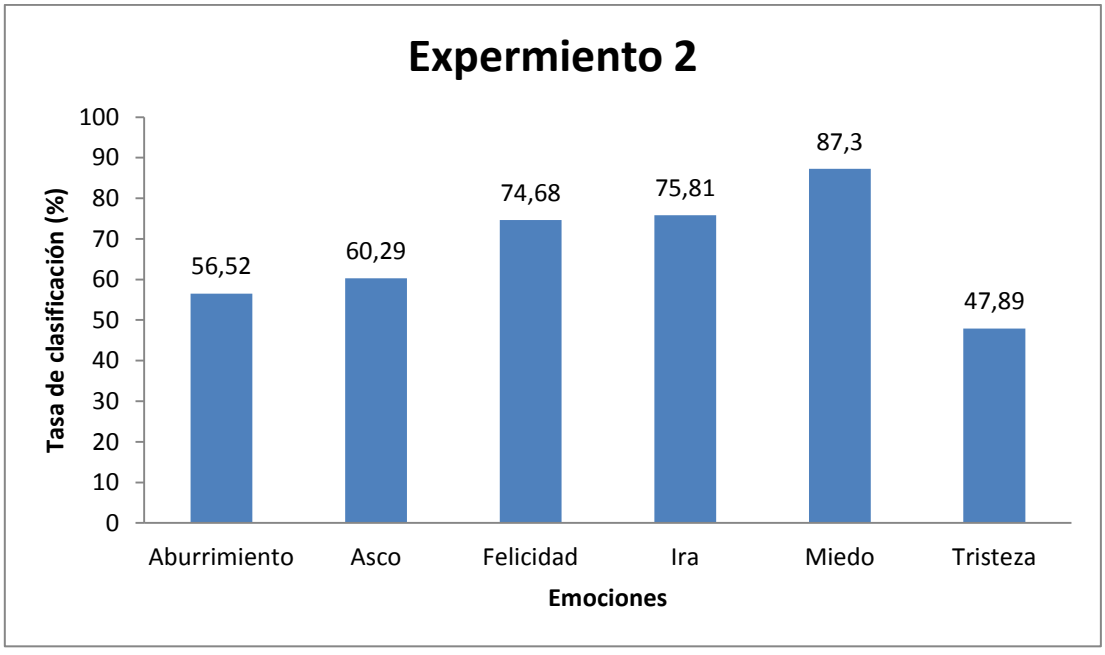

Figura 28. Tasa de clasificación promedio de las emociones del experimento 2

Observando la figura 28 se puede concluir que realizando la extracción de características espectrales a nivel de clase se detecta mejor la emoción miedo con un

87,3% de tasa de clasificación media, y que la peor emoción reconocida es la tristeza con un 47,89% de tasa de clasificación media.

4.4. Extracción de características prosódicas a nivel de expresión

Como se ha mencionado en el capítulo 2, estudios previos sobre el análisis de la emoción en el habla utilizaron diversos cálculos estadísticos de la frecuencia fundamental o Pitch (F_0) y de la frecuencia del primer formante. Por ello, en este trabajo, también se van a probar este tipo de características prosódicas para el reconocimiento automático de emociones.

Además, se van a extraer algunas características de la prosodia relacionadas con la calidad de la voz, como son el Jitter y el Shimmer, ambos conceptos explicados en el capítulo 3. Para ello, se va a utilizar el software Praat que se explica en el estudio de Boersma y Weenink de 2001 [36]. También se va a calcular la duración relativa de los segmentos sonoros de cada fichero de voz para poder caracterizar el ritmo del habla, y la energía espectral relativa por encima de los 500 Hz (HF500).

Se van a realizar algunos cálculos estadísticos a nivel de expresión, tales como valor medio, desviación estándar, mínimo y máximo de F_0, además de su primera derivada y la frecuencia del primer formante, F_1.

En total, el conjunto de características prosódicas a nivel de expresión va a contener 16 parámetros:

- Media, desviación típica, mínimo y máximo de F_0, derivada de F_0.
- Media, desviación típica, mínimo y máximo de F_1.
- Jitter, Shimmer, HF500
- Duración relativa de los segmentos sonoros de la voz.

Para calcular la frecuencia fundamental F_0 de la señal de voz se ha utilizado la función de Matlab llamada “fxrapt” contenida en el paquete “voicebox”, cuyo formato se muestra a continuación:

```
[fx,tt]=fxrapt(x,fs,'u');
```

Donde x es la señal de entrada, f_s la frecuencia fundamental y el parámetro ‘u’ permite incluir los fragmentos sordos en el vector de salida, los cuales podrán identificarse con el valor ‘NaN’ (“Not a number”).

Con el objetivo de hacer independiente la frecuencia fundamental con el género de cada locutor, los parámetros de F_0 se normalizan con la media y la desviación típica de cada locutor. Antes del proceso de parametrización en sí, para cada locutor, se calcularon los valores de media y desviación típica de la frecuencia fundamental de sus ficheros de voz y se almacenaron en otro fichero que será el que después se utilice.

Los parámetros de calidad de la señal de voz, Jitter, Shimmer y HF500 se extrajeron con el programa Praat. Cada fichero de voz tiene asociado un fichero con extensión ".jsh" donde se encuentran los valores de estos tres parámetros. Por tanto, para obtener estos parámetros hubo que leer el fichero de extensión ".jsh" asociado a cada uno de los ficheros de voz que se analizaron. Estos parámetros se calculan siempre a nivel de expresión, por tanto serán los mismos valores que se utilicen para los experimentos a nivel de clase.

La duración relativa de los segmentos sonoros de cada uno de los ficheros se calculó teniendo en cuenta la duración de la parte sonora con respecto a la duración total de cada expresión.

En la tabla 14 se muestran las diferentes pruebas realizadas para obtener los mejores resultados en este tipo de experimento.

EXPERIMENTO 3	Prueba 1	Prueba 2	Prueba 3	Prueba 4	Prueba 5	Prueba 6
Parámetros F0 (min,max,media,std)	Sí	Sí	Sí	Sí	Sí	Sí
Derivada F0 (min,max,media,std)	No	Sí	Sí	Sí	Sí	No
Parámetros F1 (min,max,media,std)	No	No	Sí	No	No	No
Jitter, Shimmer, HF500	No	No	No	Sí	Sí	Sí
Duración	No	No	No	No	Sí	Sí
BAC promedio (%)	49,24	50,89	31,57	52,73	**60,32**	59,37
Tasa de clasificación promedio (%)	55,19	56,51	35,1	55,85	63,13	62,47

Tabla 14. Pruebas del experimento 3

En este caso, se puede observar que el mejor experimento es el resultante de extraer la media, la desviación típica, el mínimo y el máximo de F_0; el mínimo, máximo, media y desviación típica de la derivada de F_0; los parámetros de calidad Jitter, Shimmer y HF500; y la duración relativa de la señal de voz.

En la tercera prueba se llegó a la conclusión de que la extracción del primer formante de la señal de voz no mejora los resultados y por eso se descartó en las siguientes pruebas.

La matriz de confusión resultante en la prueba 5 se muestra en la tabla 15.

Emoción interpretada	Emoción identificada					
	Felicidad	Tristeza	Ira	Miedo	Asco	Aburrimiento
Felicidad	61	1	11	-	7	15
Tristeza	-	32	-	21	8	4
Ira	6	-	50	-	1	1
Miedo	-	27	-	93	13	1
Asco	1	6	-	11	33	8
Aburrimiento	10	5	1	2	7	17
Total	79	71	62	127	69	46

Tabla 15. Matriz de confusión de la mejor prueba del experimento 3

Comparando los resultados, se puede apreciar que esta técnica confunde muy a menudo la emoción tristeza con la emoción miedo, ya que el número de ficheros bien reconocidos es 32 y el número de ficheros confundidos es de 27.

En la figura 29 se muestran las tasas de clasificación promedio de cada una de las clases:

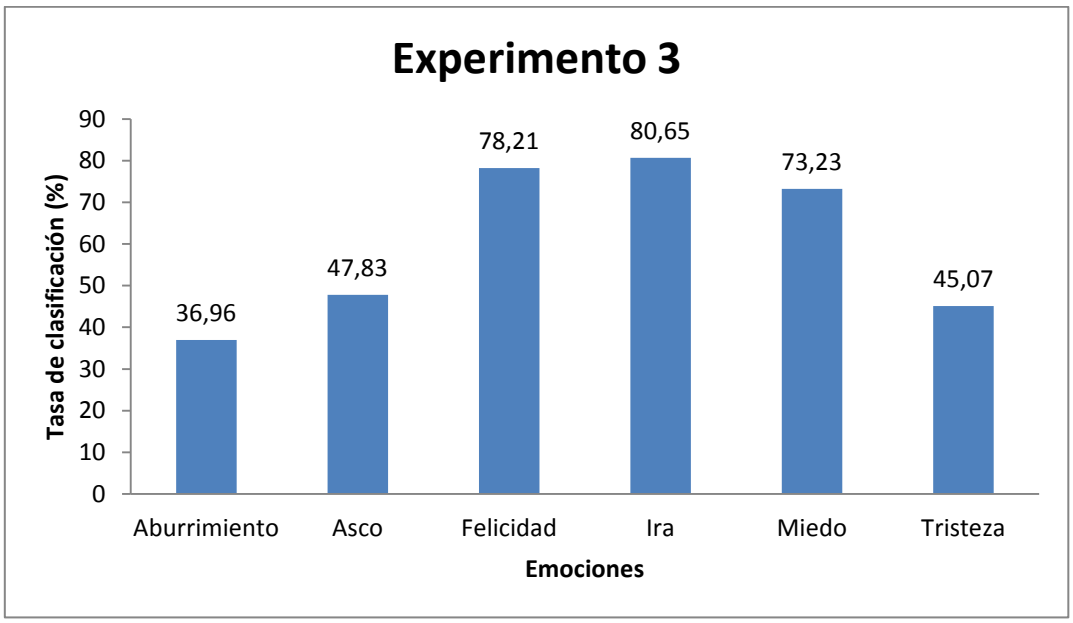

Figura 29. Tasa de clasificación promedio de las emociones del experimento 3

Si se observa la figura 29 se puede concluir que en este experimento la emoción mejor reconocida es la ira con una tasa de clasificación promedia de 80,65%, mientras que la peor reconocida es el aburrimiento con un 36,96%.

4.5. Extracción de características prosódicas a nivel de clase

En lugar de calcular estadísticos a nivel de expresión como en el apartado anterior, las características prosódicas a nivel de clase corresponden a los estadísticos de los valores de F0 y primer formante de las vocales átonas y las vocales acentuadas. No se calculan estos parámetros sobre la clase de las consonantes debido a que no están definidos para fonemas sordos.

Sin embargo, Jitter, Shimmer y HF500 se calculan sobre la expresión entera, igual que en el apartado anterior, ya que son parámetros que miden la calidad de la voz. El conjunto de características prosódicas de nivel de clase se compone de 36 elementos individuales:

- Media, desviación típica, mínimo y máximo de F_0, derivada de F_0 sobre las vocales átonas.
- Media, desviación típica, mínimo y máximo de F_0, derivada de F_0 sobre las vocales acentuadas.
- Media, desviación típica, mínimo y máximo de F_1 sobre las vocales átonas.
- Media, desviación típica, mínimo y máximo de F_1 sobre las vocales acentuadas.
- Jitter, Shimmer, HF500
- Duración relativa de los segmentos de la voz.

Este experimento no se ha llegado a implementar debido a que en estudios anteriores se ha demostrado que la extracción de este tipo de parámetros no supone ninguna mejora en las tasas de clasificación [2].

4.6. Extracción de características combinadas

Con el fin de investigar el funcionamiento de las características espectrales en combinación con las características prosódicas, se creó una parametrización combinada formada por la concatenación de las características espectrales a nivel de clase y las características prosódicas a nivel de expresión.

Se han realizado diferentes pruebas combinando los mejores resultados obtenidos en los experimentos anteriores y cuyos resultados se muestran en la tabla 16.

EXPERIMENTO 5	Prueba 1	Prueba 2	Prueba 3
E --> Log-Energy	Sí	Sí	Sí
0 --> Coeficiente de orden 0	No	No	No
d --> Coeficiente delta	Sí	Sí	Sí
D --> Coeficiente delta-delta	Sí	Sí	Sí
Media	Sí	Sí	Sí
Desviación Típica	Sí	Sí	Sí
Skewness	No	Sí	Sí
Parámetros F0 (min,max,media,std)	Sí	Sí	Sí
Derivada F0 (min,max,media,std)	Sí	Sí	Sí
Parámetros F1 (min,max,media,std)	No	No	No
Jitter, Shimmer, HF500	Sí	Sí	Sí
Duración	Sí	Sí	No
BAC promedio (%)	69,24	68,43	68,43
Tasa de clasificación promedio (%)	70,8	71,02	71,02

Tabla 16. Pruebas del experimento 5

El mejor resultado con un BAC del 69,24% se ha obtenido en la prueba 1, donde se han combinado las características que mejores resultados han dado en los experimentos 2 y 3. En este caso, el conjunto combinado consta de 90 características.

Las pruebas 2 y 3 son exactamente iguales exceptuando que en la prueba 3 no se usa la duración de los sonidos sonoros. Se puede comprobar que éste parámetro no influye en este experimento, ya que los resultados obtenidos son idénticos para ambas pruebas.

A continuación, en la tabla 17 se muestra la matriz de confusión resultante en la prueba 1.

Emoción interpretada	Emoción identificada					
	Felicidad	Tristeza	Ira	Miedo	Asco	Aburrimiento
Felicidad	63	-	5	-	9	1
Tristeza	1	40	-	21	6	10
Ira	5	-	48	-	1	-
Miedo	-	20	-	99	7	4
Asco	7	7	9	3	41	2
Aburrimiento	3	4	0	3	4	29
Total	79	71	62	126	68	46

Tabla 17. Matriz de confusión de la mejor prueba del experimento 5

En la tabla 17 se puede apreciar que la emoción tristeza se reconoce de una forma satisfactoria en 40 ocasiones, pero que se confunde con la emoción miedo 20 veces. Sin embargo, la emoción miedo se reconoce correctamente en 99 ocasiones frente a los 126 ficheros que se tienen de esta categoría.

En la figura 30 se comparan las tasas de clasificación promedio de cada una de las clases obtenidas en la prueba 1.

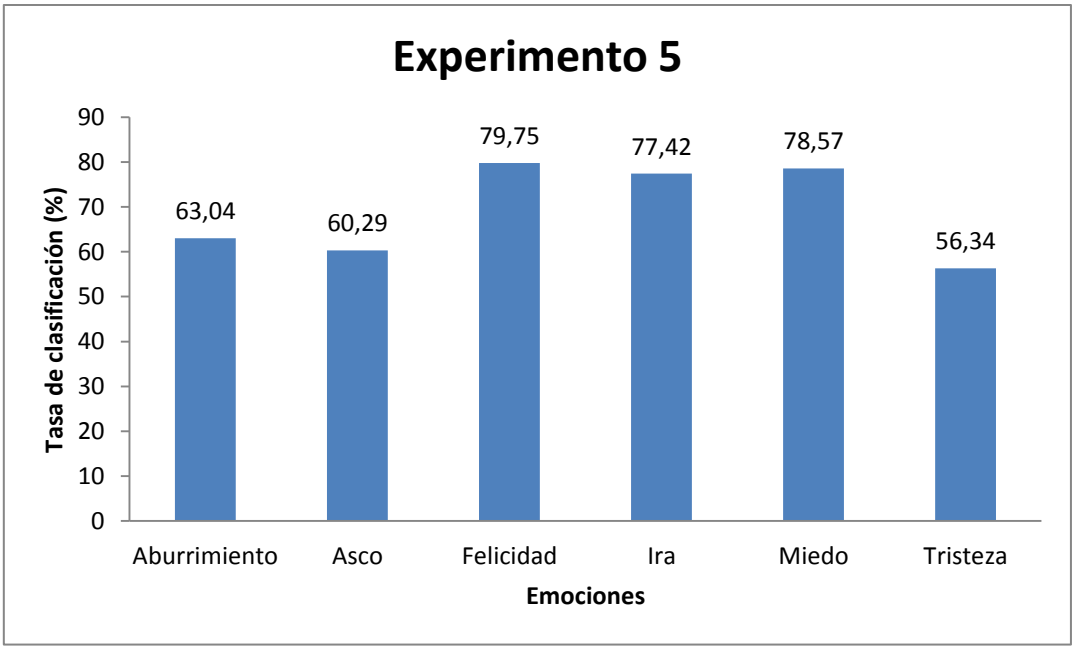

Figura 30. Tasas de clasificación promedio de las emociones del experimento 5

En la figura 30 se observa que la emoción que mejor se reconoce es la felicidad con una tasa de clasificación promedio de 79.75%, siendo el asco la emoción peor reconocida con una tasa del 60.29%.

4.7. Conclusiones de los experimentos

El objetivo principal de éste trabajo era realizar experimentos similares a los que se desarrollaron en el estudio [2] y utilizando técnicas similares igualar los resultados obtenidos o incluso mejorarlos.

En la tabla 18 se comparan los resultados obtenidos en ambos experimentos:

Características	BAC (%) Proyecto	Bac (%) Estudio
Características espectrales a nivel de expresión	64,44	67
Características espectrales a nivel de clase	67,08	75,9
Características prosódicas a nivel de expresión	60,32	68,1
Características prosódicas a nivel de clase	-	68,6
Características combinadas	69,24	78,2

Tabla 18. Comparación de resultados

Como se puede observar en la tabla 18 los resultados obtenidos en este proyecto se acercan a los resultados del estudio [2], pero no se consiguen mejorar.

En el caso de las características espectrales a nivel de expresión, esta diferencia en los resultados puede ser debida a la forma de calcular los coeficientes cepstrales MFCC, dado que en el artículo de referencia no se mencionan ciertos parámetros importantes de configuración (como, por ejemplo, si se ha aplicado algún tipo de normalización), ni se indica el software con el que se han extraído los coeficientes. En el caso de las características espectrales a nivel de clase, además del motivo mencionado anteriormente, hay que destacar que la segmentación en fonemas disponible en la base de datos contenía algunos errores y desconocemos si en el artículo se resolvieron manualmente o se eliminaron los ficheros conflictivos.

En la extracción de las características prosódicas se observa una mayor diferencia entre los resultados. De nuevo en el artículo, no se mencionan algunos parámetros de configuración importantes para la extracción de estas características (por ejemplo, no se indica si se utiliza una frecuencia fundamental interpolada para las tramas sordas, se descartan dichas tramas o se les asigna una frecuencia fundamental nula). En particular, en este proyecto, el cálculo del primer formante no ha aportado ninguna mejora al experimento, por lo que se puede concluir que la técnica utilizada para el cálculo de estas características puede no haber sido la más correcta.

En la figura 31 se muestran los mejores resultados obtenidos en los experimentos de este trabajo.

Se observa que el mejor resultado con un BAC del 69.24% se ha obtenido en la combinación de características, mientras que el peor, con un resultado del 60.32% se ha obtenido en la extracción de características prosódicas a nivel de expresión.

En cualquier caso, los resultados obtenidos son razonables, aunque sería conveniente probar nuevas alternativas de extracción de parámetros para conseguir unas tasas de reconocimiento más próximas a las del estudio de referencia.

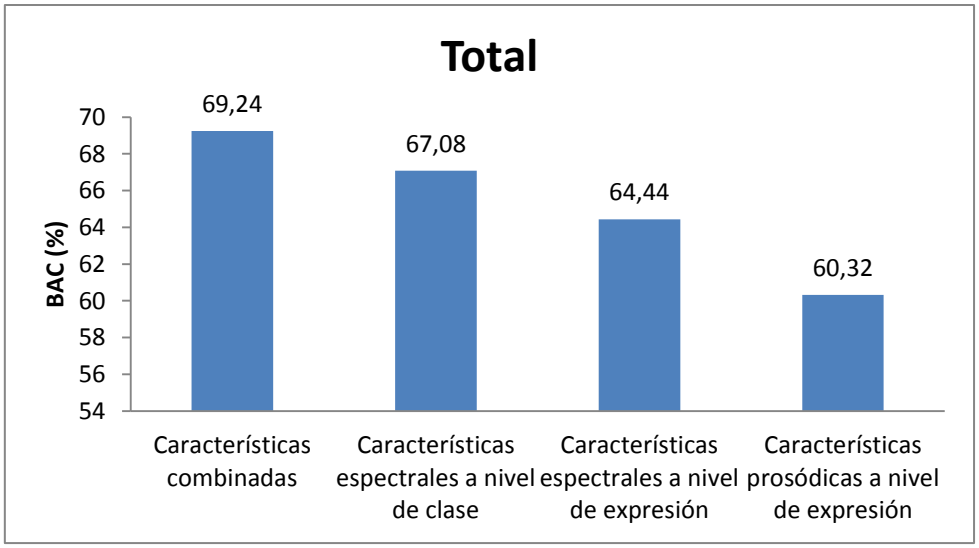

Figura 31. Resumen de los resultados de los experimentos

Capítulo 5

Gestión del proyecto

En este último capítulo se presenta una estimación de los recursos económicos y humanos para la realización de este trabajo fin de grado. Primero se hará una descripción de las fases de trabajo para después comentar qué recursos han sido necesarios.

5.1. Introducción

Para el desarrollo de cualquier proyecto es de mucha utilidad realizar una planificación para poder estimar los costes y su duración. Es necesario tener en cuenta todos los factores que pueden alterar esa planificación, tanto en los costes, como en el ámbito temporal [37].

Lo primero de todo es definir los objetivos del proyecto, es decir, las funcionalidades que se van a investigar o desarrollar, para después diferenciarlas en distintas fases de trabajo y organizarlas temporalmente en un calendario. Si esta planificación se hace de una forma correcta, se pueden definir los recursos que se van a emplear desde el primer momento, lo que supondrá un ahorro en los costes. Además, siguiendo el calendario, se puede conocer en qué punto de desarrollo se encuentra el proyecto en cada momento [38].

Tener una metodología de trabajo, permite encontrar errores para futuros proyectos y así aprender de la experiencia. Un proyecto correctamente gestionado podría seguir el siguiente esquema de la figura 32:

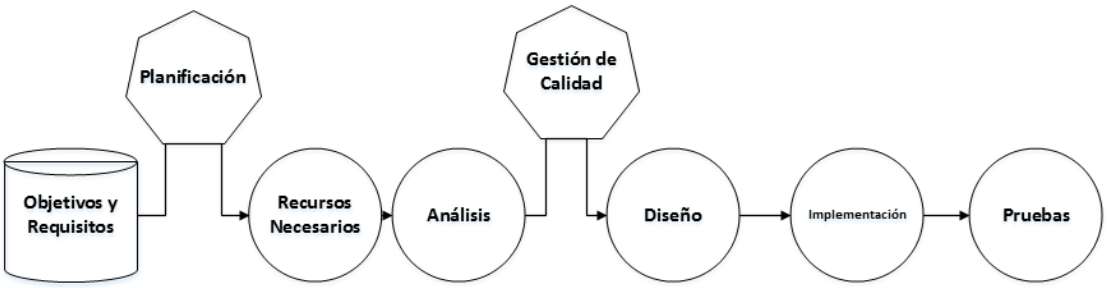

Figura 32. Esquema de una metodología de trabajo

5.2. Fases de trabajo

Para la realización de este trabajo se han definido las siguientes fases:

- Fase 1: **Documentación**

El primer paso es hacer una búsqueda de información sobre el estado del arte de reconocimiento de emociones, leer trabajos previos de otros autores, familiarizarse con las técnicas empleadas, y comprender los conceptos básicos para el entendimiento del trabajo que se va a desarrollar.

El objetivo de esta primera fase es adquirir una base, tanto de conceptos teóricos, como de familiarización con la herramienta Matlab, que es la que se ha usado para implementar el código del programa.

Para la realización satisfactoria de esta primera fase, ha sido necesario un tiempo aproximado de 100 horas.

- Fase 2: **Desarrollo del Software**

Esta fase consiste en la implementación de un software diferenciado en cinco módulos distintos, correspondiendo cada uno a los cinco tipos de características que se han extraído a partir de los ficheros de voz pertenecientes a la base de datos “Berlin

Data Base". Previo paso a la implementación de cada uno de los módulos, ha sido necesario hacer un estudio sobre las características necesarias para extraer en cada uno de ellos.

El tiempo estimado para cada módulo ha sido:

- Módulo 1: 50 horas
- Módulo 2: 120 horas
- Módulo 3: 80 horas
- Módulo 5: 50 horas

En total se estima que han sido empleadas 300 horas para la realización de esta fase.

- Fase 3: **Experimentos**

El número de experimentos realizados no es el mismo para cada uno de los módulos. Dependiendo de las características que se han extraído, se han podido modificar unos parámetros u otros. Además, el tiempo de proceso de cada experimento es diferente debido a los diferentes parámetros que se extraen en cada módulo.

El tiempo estimado para cada módulo ha sido:

- Módulo 1: 10 horas
- Módulo 2: 15 horas
- Módulo 3: 30 horas
- Módulo 5: 15 horas

En total se estima que han sido empleadas 70 horas para la realización de esta fase.

- Fase 4: **Redacción de la memoria**

Dentro de esta fase, se puede hacer una diferenciación de 4 partes:

- Parte 1: Organización y estructura de la memoria. Se realiza un índice con el orden de los contenidos que se van a desarrollar.
- Parte 2: Recopilación de información. Buscar todas las referencias bibliográficas que puedan ser necesarias para añadir información útil a los contenidos.
- Parte 3: Redacción de cada uno de los capítulos de la memoria.

- Parte 4: Revisión de la memoria. Estructurar y referenciar organizadamente los contenidos, redactar el resumen y el abstract, elaboración de la portada y los agradecimientos.

El tiempo para la realización de esta fase se estima en 200 horas.

- Fase 5: **Presentación del trabajo de fin de grado**

Para la realización de esta fase es necesario saber escoger la información más relevante del trabajo, para poder exponerla de manera organizada y amena. Se estima que se han invertido 20 horas en la realización de esta fase.

En la tabla 19 se puede ver un resumen del tiempo estimado para la realización de cada una de las fases de trabajo:

Nombre de la Tarea	Inicio	Fin	Duración
Documentación	01/10/2013	25/11/2013	100 horas
Desarrollo del Software	01/12/2013	15/05/2014	400 horas
Módulo 1	01/12/2013	25/12/2013	50 horas
Módulo 2	08/01/2014	10/03/2014	120 horas
Módulo 3	12/03/2014	02/04/2014	80 horas
Módulo 5	12/04/2014	10/05/2014	50 horas
Experimentos	10/05/2014	31/05/2014	100 horas
Memoria	10/01/2014	15/06/2014	200 horas
Organización y estructura	10/01/2014	25/01/2014	20 horas
Recopilación de Información	25/01/2014	20/02/2014	40 horas
Redacción de la memoria	23/02/2014	01/06/2014	120 horas
Revisión de la memoria	01/06/2014	15/06/2014	20 horas
Presentación del TFG	16/06/2014	01/07/2014	20 horas

Tabla 19. Fases de trabajo

En la figura 33 se muestra el diagrama de Gantt con la planificación del proyecto. El diagrama de Gantt se ha realizado con el software de la empresa Gantt Project, siguiendo las instrucciones que se indican en el tutorial que muestran en [E17].

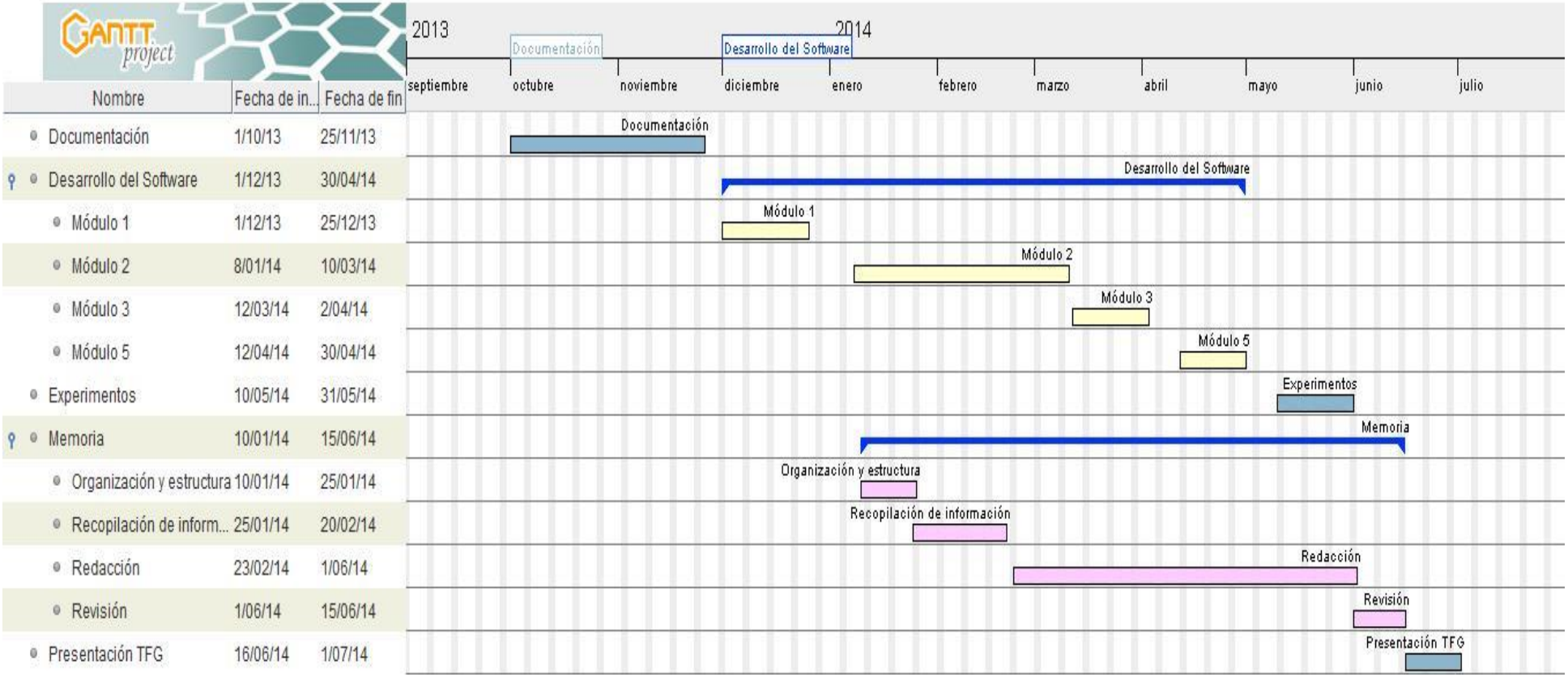

Figura 33. Diagrama de Gantt

5.3. Recursos de trabajo

Los recursos utilizados durante la realización de este trabajo se pueden dividir en recursos humanos y recursos materiales. A continuación se hará un estudio de los costes totales debido al desarrollo del trabajo. Para la estimación del presupuesto se ha utilizado la plantilla proporcionada por la Universidad Carlos III de Madrid [E18].

5.3.1. Recursos humanos

Para la estimación de los costes atribuidos a los recursos humanos es necesario detallar los perfiles profesionales de las personas que han participado en este trabajo.

- **Tutor del proyecto**: Profesional titulada en ingeniería de telecomunicaciones y doctora ingeniera de telecomunicaciones. Tiene experiencia en el sector del reconocimiento automático del habla y capacidad para liderar, dirigir y solventar problemas. Adquiere el papel de Ingeniero Senior.
- **Investigador:** Alumna cursando el Grado en Ingeniería de Sistemas Audiovisuales a falta de presentar el trabajo de fin de grado. Tiene conocimientos de señales, de tratamiento digital de audio, así como de desarrollo software.

En la tabla 20 se muestra el desglose de costes asociados a cada persona.

Apellidos y nombre	Categoría	Dedicación (mes)	Coste hombre (mes)	Coste (Euro)
Gallardo Antolín, Ascensión	Ingeniero Senior	1*	4.289,54	4.289,54
Navidad Peñalba, Irene María	Ingeniero	9	2.694,39	24.249,51
	Hombres mes 9		**Total**	**28.539,05 €**

* El mes de trabajo está dividido en reuniones semanales de 1h aproximadamente

Tabla 20. Costes personales

5.3.2. Recursos materiales

En la tabla 23 se muestran los recursos materiales, tanto software como hardware, necesarios para la realización de este trabajo, así como sus costes asociados.

Descripción	Coste (Euro)	% Uso dedicado proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable
Laptop Acer Aspire 5755G	589,00	100	9	60	88,35
Matlab Home License	105,00	100	9	60	15,75
Office 2013 Home & Student PKC	99,99	100	9	60	15,00
Visio 2013	739,00	100	9	60	110,85
Gantt Project	0,00	100	9	60	0,00
VLC Media Player	0,00	100	9	60	0,00
				Total	**229,95 €**

Tabla 21. Costes materiales

Los costes del software y del hardware utilizado se han presupuestado suponiendo que el uso de cada uno de ellos es del 100%. Los precios que se muestran son sin IVA.

El coste imputable, E, de cada elemento se calcula a través de la fórmula de la amortización, ecuación 30.

$$E = \frac{A}{B} * C * D$$

Ecuación 30. Fórmula de la amortización

Dónde: A, nº de meses desde la fecha de facturación en que el equipo es utilizado.

B, Periodo de deprecación (60 meses).

C, Coste (sin IVA)

D, % dedicado al proyecto (Por lo general, siempre es el 100%).

5.4. Costes totales

Para calcular los costes totales que suponen la realización de este trabajo, es necesario tener en cuenta los costes indirectos, que por lo general son del 20%. Estos costes indirectos pueden ser debidos a viajes de negocios, imprevistos o mejoras de última hora.

Considerando que el proyecto se factura en España, se tiene que aplicar el 21% de IVA sobre la cantidad total.

En la tabla 24 se puede ver un desglose de los conceptos con sus costes asociados.

Concepto	**Costes Totales**
Personal	28.539
Amortización	230
Costes Indirectos	5.754
Costes Totales sin IVA	34.523
Total	**41.773 €**

Tabla 22. Costes totales

El coste total asociado al desarrollo del proyecto asciende a 41.773 € (cuarenta y un mil setecientos setenta y tres euros).

Conclusiones y líneas futuras

En este trabajo fin de grado se ha desarrollado una aplicación en Matlab que consiste en un sistema de reconocimiento automático de emociones a partir de la voz de un hablante. El sistema que se ha desarrollado consta de dos etapas: parametrizador y clasificador.

En la etapa del parametrizador se han extraído características espectrales y prosódicas y se han calculado sus estadísticos a nivel de clase y a nivel de expresión. En relación con las características espectrales se han extraído los coeficientes mel-cepstrales (MFCC) junto con sus parámetros Delta-MFCC y Delta-Delta-MFCC. Como características prosódicas se han extraído la frecuencia fundamental, la frecuencia del primer formante, la duración relativa de los segmentos sonoros y algunos parámetros de calidad de la señal de voz como son jitter, shimmer y la energía espectral relativa por encima de los 500Hz (HF500).

En total se han implementado cuatro módulos distintos de extracción de características:

- Características espectrales a nivel de expresión
- Características espectrales a nivel de clase
- Características prosódicas a nivel de expresión
- Características combinadas

En la etapa del clasificador se han comparado las características extraídas con los patrones de emociones que se han obtenido a partir de una base de datos previamente etiquetada mediante un proceso de entrenamiento. La técnica de clasificación que se ha usado para realizar estas comparaciones está basada en máquinas de vector soporte.

Para cada uno de los conjuntos de características antes mencionados se han realizado diversas pruebas modificando algunos de los parámetros. Aunque los resultados obtenidos no mejoran a los del estudio [2], tomado de referencia, se puede concluir que las características espectrales a nivel de clase aportan más información sobre las emociones que las características prosódicas. Sin embargo, los mejores resultados se obtienen combinando las características espectrales a nivel de clase con las características prosódicas a nivel de expresión.

En general se puede decir que los objetivos inicialmente planteados se han cumplido, ya que la aspiración principal era estudiar los sistemas automáticos de reconocimiento de emociones en la voz y adquirir una base sólida para poder implementar una aplicación que permitiera evaluar las técnicas aprendidas.

Aunque los objetivos planteados inicialmente se han cumplido, a partir de los resultados experimentales obtenidos y de las principales conclusiones extraídas en este proyecto, se plantean diferentes líneas de investigación en el campo del reconocimiento de las emociones en el habla, que a continuación se detallan.

El primer paso sería mejorar el sistema de extracción de algunas de las características, como por ejemplo el primer formante, que en este caso no ha aportado ninguna mejora. También se podrían añadir más características como puede ser la intensidad de la voz.

Para poder observar la eficacia del sistema implementado, se podrían utilizar diferentes bases de datos, grabadas en diferentes idiomas y condiciones. Así se podría estudiar la influencia de los idiomas en este tipo de sistemas automáticos de reconocimiento de emociones.

También se podrían utilizar y combinar otras técnicas de clasificación como pueden ser los modelos ocultos de Markov o los modelos de mezclas de gaussianas. De esta forma se podrían medir las prestaciones de cada una de estas técnicas a la hora de clasificar emociones.

Por otra parte, sería interesante implementar una aplicación on-line de este sistema automático de reconocimiento de emociones. Dicha aplicación debería grabar un fragmento de voz, limpiar el ruido de fondo (si lo hubiera) aplicando alguna técnica de mejora de voz y extraer ciertas características de ese fragmento de voz para que, a partir de ellas, se detectara en un plazo corto de tiempo, cómo se encuentra emocionalmente el hablante.

Otra línea de interés sería la mejora de este tipo de sistemas mediante la combinación de descriptores de audio y video. Así se podrían extraer tanto características de la voz como características de las expresiones faciales según la emoción, construyendo de ésta manera un sistema mucho más completo y robusto.

Otra mejora importante que podría realizarse sería aumentar el número de emociones que pueden identificarse.

Utilizando las técnicas presentadas en este proyecto podrían implementarse algunas aplicaciones como las que se exponen a continuación:

- Reproductor de música que genere una lista de canciones acorde al estado de ánimo del usuario y que podrá detectar cuando éste diga una frase al inicio de la sesión. Esta aplicación se podría aplicar también al mundo del cine, donde se recomienden películas de acuerdo al estado emocional del usuario.
- Los sistemas automáticos de reconocimiento de emociones podrían ser muy útiles si se aplican en el campo de la psicología, donde podrían ayudar a detectar depresiones en pacientes o incluso algunas enfermedades como el autismo.

Conclusions and future lines

In this final degree project, an application has been created in Matlab based on an automatic recognition system of emotions from the voice of a speaker. The system developed consists of two stages: parameter assignment and classifier.

In the stage of parameter assignment, prosodic and spectral features have been extracted and their statistics have been calculated at class and expression level. Regarding the spectral features, coefficients mel-cepstral (MFCC) were extracted along with its parameters Delta-MFCC and Delta-Delta-MFCC. Fundamental frequency, the frequency of the first formant, the relative duration of the sound segments, and quality parameters of the voice signal such as jitter and shimmer relative spectral energy above 500 Hz (HF500), have been extracted like prosodic features.

Four different modules of feature extraction have been implemented in total:

- Utterance-level spectral features
- Class level spectral features
- Utterance-level prosodic features
- Combined features

In the classifier step, features extracted with emotion patterns that have been obtained from a database previously labeled by a training process, have been compared. The classification method used to perform these comparisons is based on support vector machines.

For each of the sets of features above before mentioned, several tests have been made changing some parameters. Although the results obtained do not improve the study [2], taken by reference, it can be concluded that the class-level spectral features provide more information about the emotions than prosodic features. However, the best results are obtained by combining the spectral characteristics at the class level with the utterance-level prosodic features .

Based on the original objectives proposed, it could be said that they have been reached, since the main aim was to study automatic emotion recognition systems of voice, and achieve a solid foundation in order to implement an application that allows evaluating the techniques learned.

Although the targets initially planned were reached, considering the experimental results and the main conclusions obtained in this project, different research lines are proposed in the field of emotion recognition in voice signal, which are detailed below.

The first step would be to improve the extraction system of some features, such as the first formant, which in this case has provided no improvement. Besides, more features, such as the intensity of the voice, could be added.

To observe the efficiency of the implemented system, different databases recorded in different languages and conditions could be used. This research could give the possibility to study the influence of languages in this type of automatic emotion recognition.

Other classification techniques like hidden Markov models or Gaussian mixture models could be used and combined. As a result, measuring the performance of each of these techniques at the time of classifying emotions could be possible.

On the other hand, it would be interesting to implement an on-line application of the automatic emotion recognition. This application should record a voice fragment, clean background noise (if any) by applying a technique to improve voice, and extract certain features of that fragment of voice in order to detect in a short time how the speaker feels emotionally.

Another line of interest would be to improve such systems by combining video and audio descriptors, so both voice characteristics and features of facial expressions could be extracted considering the emotion, building this way a more complete and robust system.

In fact, another important improvement that could be done would be to increase the number of emotions that can be identified.

Using the techniques presented in this project, some applications could be implemented, like for example, the ones listed below:

- A music player that generates a playlist according to the user's mood and detects when he says a phrase at the beginning of the session. This application could also be applied to the world of cinema, where films could be recommended according to the user´s emotional state.
- Automatic emotion recognition systems could be very useful if we applied them in the field of psychology, where it could be helpful at the time of detecting depression in patients or even some illnesses like autism.

Bibliografía

[1]: Jesús Bernal Bermúdez, Jesús Bobadilla Sancho, Pedro Gómez Vilda. Reconocimiento de voz y de fonética acústica. Enero, 2000.

[2]: Bitouk, D., Verma, R., Nenkova, A., "Class-level spectral features for emotion recognition". Volume 52, Issues 7‐8, July‐August 2010, Pages 613‐625.

[3]: Cowie,R., Douglas -Cowie,E., Tsapatsoulis,N., Votsis,G., Kollias,S. ,Fellenz,W., Taylor,J.G; Emotion recognition in human-computer interaction in IEEE Signal Processing Magazine, Vol. 18, Issue1,pp.32-80,January 2001.

[4]: Tin Lay New, Say Wei Foo, Liyanage C. de Silva; Speech emotion recognition using hidden Markov models in Elsevier Speech Communications Journal Vol. 41, Issue 4, pp. 603-623, November 2003.

[5]: E. Kramer. E. C. Beier and A. J. Zautra. D. C. Albas, K. W. Mccluskey, and C. A. Albas.K. W. Mccluskey, D. C. Albas, and R. R. Niemi. R. V. Bezooijen, Characteristics an recognizability of vocal expressions of emotion. Dordrecht: Foris Publications, 1984.

[6]: Carroll E. Izard (Author of The Psychology of Emotions), Innate and universal facial expressions: evidence from developmental and cross-cultural research. 1994, Vol.115, Nº2, pp. 288-299.

[7]: C. Pereira, "Dimensions of emotional meaning in speech", In SpeechEmotion-2000,pp. 25-28.

[8]: M Schröder. Emotional speech synthesis: a review. INTERSPEECH, 561-564.

[9]: Spinoza´s "Ethics", An Introduction. Author: Steven Nadler. May 2006

[10]: Eriksson, E.J., Rodman, R.D., & Hubal, R.C. (2007). Emotions in speech: Juristic implications. In Muller, C. (Ed.),*Speaker Classification I: Fundamentals, Features, and Methods. Lecture Notes in Artificial Intelligence Series.* (pp. 152-173). Heidelberg, Germany: Springer Berlin".

[11]: E. S. y F Núñez y P Cortes y C Suárez, "Índice de incapacidad vocal: factores predictivos," Acta Otorrinolaringólogo, vol. 57, pp. 101–108, 2006.

[12]: X. Huang, Spoken language processing. Prentice-Hall Inc, 2001.

[13]: Carl E. Williams and Kenneth N. Stevens, "Emotions and speech: Some acoustical correlates", The Journal of the Acoustical Society of America, vol. 52, no. 4B, pp. 1238-1250,1972.

[14]: Samal, A., Iyengar, P., 1992. Automatic recognition and analysis of human faces and facial expression: a survey. Pattern Recognition 25 (1), 65–77.

[15]: Picard, R., 1997. Affective Computing. MIT Press, Cambridge, MA.

[16]: Polzin, T., Waibel, A., 2000. Emotion-sensitive Human–computer interface. In: Proceedings of the ISCA Workshop on Speech and Emotion.

[17]: Whiteside, S.P., 1998. Simulated emotions : an acoustic study of voice and perturbation measures. In: Proceedings of ICSLP 1998, pp. 699–703.

[18]: McGilloway, S., Cowie, R., Cowie, E.D., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion.

[19]: Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods.

[20]: F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech".

[21]: Konstantin Stanislavski, "El arte escénico". Siglo XXI, 2009.

[22]: Laver, J., "The Phonetic Description of Voice Quality", Cambridge University Press, Cambridge, 1980

[23]: Isaac Martín de Diego, Ángel Serrano, Cristina Conde, Enrique cabello. Técnicas de reconocimiento automático de emociones. Revista electrónica teoría de la educación. Educación y cultura en la sociedad de la información. Vol.7, Nº2, Diciembre 2006. Pags. 110-122.

[24]: Ferguson, J. Hidden Markov Models for Speech. IDA, Princeton, NJ. 1980.

[25]: Huang, X. D., Ariki, Y., y Jack, M. A. Hidden Markov Models for Speech Recognition. Edinburgh University Press. 1990.

[26]: V.N. Vapnik. Statistical Learning Theory. John Wiley & Sons, Inc, 1998.

[27]: Gustavo A. Betancourt. Las máquinas de soporte vectorial (SVMs). Scientia et Technica Año XI, No 27, Abril 2005

[28]: Emotion in Speech: Recognition and applications to call centers. Autor: Valery A. Petrushin.

[29]: Zhihong Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, \A survey of affect recognition methods: Audio, visual, and spontaneous expressions", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 1, pp. 39{58, 2009.

[30]: Carlos Ortego Resa, Detección de emociones en voz espontánea. Universidad autónoma de Madrid, Julio 2009.

[31]: C. Ferrer, M.E. Hernández-Díaz. Relación entre la frecuencia fundamental y la percepción subjetiva del Pitch. Memorias II congreso Latinoamericano de Ingeniería Biomédica, La Habana, Cuba, 23-25 Mayo 2001.

[32]: J.H.L. Hansen, \Evaluation of acoustic correlates of speech under stress for robust speech recognition", Mar 1989, pp. 31-32.

[33]: Carlos Monzo, Ignasi Iriondo y Elisa Martínez. Procedimiento para la medida y la modificación del jitter y el shimmer aplicado a la síntesis del habla expresiva. V Jornadas en Tecnología del Habla.

[34]: J.H.L. Hansen and S. Patil, \Speech under stress: Analysis, modeling and recognition", in Speaker Classi_cation (1). 2007, vol. 4343 of Lecture Notes in Computer Science, pp. 108-137, Springer.

[35]: Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. Department of Computer Science. National Taiwan University, Taipei, Taiwan.

[36]: Boersma, P., Weenink, D., 2001. Praat, a system for doing phonetics by computer. Glot Internat., 341–345.

[37]: W.S. Humphrey: “A discipline for software engineering”. Ed. Adisson Wesley.1995.

[38]: A. Cuevas, (2003), Gestión del proceso software, Ed: Editorial Universitaria Ramón Areces

Enlaces virtuales

[E1]: "Inteligencia emocional," Tech. Rep., http://www.inteligencia-emocional.org (Último acceso: Junio 2014).

[E2]: http://www.ventanaskline.es/ventanas-de-aluminio/ventanas-aluminio/solucion-ruido (Último acceso: Junio 2014).

[E3]: http://recursos.cnice.mec.es/media/radio/bloque2/pag5b.htm (Último acceso: Junio 2014).

[E4]: http://www.rinconpsicologia.com/2010/11/la-respiracion-diafragmatica-una.html (Último acceso: Junio 2014).

[E5]: http://www.cancer.gov/espanol/pdq/tratamiento/laringe/Patient/page1 (Último acceso: Junio 2014).

[E6]: http://paginaspersonales.deusto.es/airibar/Fonetica/Apuntes/02.html (Último acceso: Junio 2014).

[E7]: http://www.fceia.unr.edu.ar/prodivoz/Modelo_Produccion_Voz_bw.pdf (Último acceso: Junio 2014).

[E8]: Berlín Data Base, http://www.expressive-speech.net/emoDB/ (Último acceso: Junio 2014).

[E9]: http://autismodiario.org/2013/07/08/proyecto-emociones-una-aplicacion-que-ayuda-al-desarrollo-de-la-empatia-en-los-ninos-con-autismo/ (Último acceso: Junio 2014).

[E10]: http://emotional-apps.com/#meit (Último acceso: Junio 2014).

[E11]: http://emotional-apps.com/#picfeel (Último acceso: Junio 2014).

[E12]: http://www.sciencedirect.com/science/article/pii/S0020025510003610 (Último acceso: Junio 2014).

[E13]: http://emotionsense.org/ (Último acceso: Junio 2014).

[E14]: https://itunes.apple.com/us/app/in-flow-mood-and-emotion-diary/id549101905?mt=8 (Último acceso: Junio 2014).

[E15]: http://bibing.us.es/proyectos/abreproy/12054/fichero/MEMORIA%252F8.Cap%EDtulo+3.pdf (Último acceso: Junio 2014).

[E16]: http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (Último acceso: Junio 2014).

[E17]: http://www.ganttproject.biz/ (Último acceso: Junio 2014).

[E18]: Universidad Carlos III de Madrid: "Plantilla presupuesto TFG". Spain. 2014. http://portal.uc3m.es/portal/page/portal/administracion_campus_leganes_est_cg/proyecto_fin_carrera (Último acceso: Junio 2014).