Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas
Ingeniería en Informática

Reconocimiento Automático de Emociones en Contenido Multimedia

Autores: Alvarez, Ramiro Andrés
Fadil, Carim
Director: Rufiner, Hugo Leonardo
Codirector: Martínez, Cesar

Santa Fe, 2014

A nuestras familias, sin las cuales no hubiéramos podido afrontar este desafío.

A aquellos que ven más en nosotros, y que con su mirada nos hacen mejores.

Resumen

En los últimos años se ha puesto mucho esfuerzo en la detección de emociones en el discurso y en la detección de emociones en expresiones faciales con el objeto de lograr una comunicación más "natural" y efectiva entre computadoras y seres humanos. Se han propuesto diversos métodos que usan diferentes modelos para intentar reconocer automáticamente emociones en el discurso y en las expresiones faciales. En este trabajo se presenta el diseño e implementación de un sistema multimodal de reconocimiento de emociones en contenido audiovisual, que aprovecha la correlación de la información emocional presente en ambos canales: el video capturado de personas expresando emociones y la voz. El sistema cuenta con dos módulos diferentes: extracción de características y clasificación.

Para la extracción de características se analiza la evolución temporal de la posición de marcadores faciales de cada expresión presente en un video, mientras que del audio se extraen características prosódicas y espectrales. Para probar el clasificador propuesto se utilizó la base de datos *RML Emotion Database* la cual presenta las 6 emociones universales: ira, miedo, asco, sorpresa, alegría y tristeza. Esta base de datos utiliza 10 sentencias de discurso diferentes, en 6 idiomas distintos y actuadas por 8 personas de diferentes culturas.

El clasificador utilizado es una red neuronal profunda diseñada mediante la técnica de autocodificadores profundos apilados. Se diseñaron diferentes vectores de características para probar el desempeño del clasificador. Se evaluó el desempeño del mismo frente a las distintas variantes de vectores de características y se obtuvieron resultados de clasificación cercanos al 80 % independientemente de la cultura y el idioma. Se comparó el clasificador con un perceptrón multicapa, con el que se obtiene un desempeño cercano al 78 %.

Se realizaron experimentos monomodales y se obtuvieron tasas de acierto del 77 % para audio y del 75 % para video, con lo que se concluye que el enfoque multimodal mejora el rendimiento de clasificación.

Finalmente, se realizaron diversas pruebas de clasificación monoidioma. Se evaluó el desempeño del clasificador propuesto solo para frases en el idioma inglés, para todas las variantes de los vectores de características, y se llegó a tasas de acierto similares a las pruebas multiidioma. Con esto se concluye que se diseñó un método que es robusto frente a diferentes variantes culturales e idiomas hablados.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

A medida que crece el número y la funcionalidad de las máquinas y computadoras con las que tenemos interacción en nuestra vida cotidiana, se vuelve más importante lograr una buena y fluida forma de comunicación con las mismas. Es esperable entonces que pueda haber una interacción menos estructurada entre seres humanos y computadoras/robots. Para que esto sea posible, es necesario poder diseñar sistemas inteligentes que puedan comunicarse con las personas no sólo intercambiando información lógica sino también emocional. Es en este sentido que el reconocimiento de la emoción humana se vuelve un paso muy importante. Mehrabian [66] estableció que en la transmisión de mensajes con carga emocional, las palabras sólo transmiten un 7 % de la información, mientras que la prosodia (rasgos fónicos que afectan a la métrica de la voz) contribuye un 38 % y el lenguaje corporal (las expresiones faciales, gestos, etc.) lo hacen en un 55 %. Por otro lado, la teoría evolucionista sobre el origen de las emociones sostiene que hay seis emociones principales que pueden ser reconocidas en las expresiones faciales de manera universal, y que estas expresiones no dependen de la cultura, sino que tienen origen biológico [29].

Se ha puesto mucho énfasis durante la última década en la investigación en el campo de la Interacción Humano-Computadora (HCI, del inglés Human-Computer Interaction). En este sentido, se han intentado aplicar modelos de comunicación originarios de las ramas de las Ciencias Sociales como la Ciencia de la Comunicación, las Neurociencias, la Psicología, etc. [18]. La comunicación entre personas cara a cara se da generalmente de manera simultánea a través del contenido del discurso, la prosodia y el lenguaje corporal. El contenido del discurso transmite información explícita del mensaje, la prosodia y el lenguaje corporal están más vinculados a la carga emocional del mensaje [78, 79]. Se ha demostrado que el procesamiento cerebral de la comunicación es multimodal y facilita la empatía y la comprensión del interlocutor cuando están presentes los tres canales [78]. Particularmente, se ha verificado que eliminar uno o ambos de estos canales (al neutralizar la emoción) resulta en una mayor dificultad de empatía y comprensión, esenciales para una comunicación efectiva. De esta manera, para poder desarrollar sistemas inteligentes que interactúen más naturalmente con los seres humanos, se debería implementar el reconocimiento de emociones humanas a partir de las expresiones faciales y de la señal de voz.

El diseño y desarrollo de sistemas automáticos de reconocimiento de emociones presenta distintos problemas:

1. Detección robusta y automática de rostros para segmentar la imagen.
2. Extracción de información de relevancia sobre las expresiones faciales que permita distinguir emociones.
3. Extracción de características en la señal de voz para obtener la información de interés que permita discernir entre las distintas emociones.
4. Reducción de dimensionalidad de la información para que pueda reducir los tiempos de cómputo y/o reducir la complejidad en el diseño del clasificador.
5. Diseño de un clasificador que aprenda los modelos subyacentes de las expresiones faciales y la voz, permitiendo así aprovechar ambos canales de información.

Por tanto, en el desarrollo de este capítulo se hará un repaso sobre el concepto de emoción y los diferentes enfoques en cuanto a la clasificación de las mismas. Se explicará cómo se relacionan las emociones con las expresiones faciales y la voz desde un punto de vista psicológico. Luego se dará una breve explicación de cómo el aparato fonador genera la señal de voz y las distintas características que ésta presenta. Finalmente se presentarán la justificación y los objetivos de este trabajo.

1.1. El Concepto de Emoción

Es necesario conceptualizar qué es lo que se entiende por emoción o estado emocional. En realidad, no hay una definición teórica establecida y aceptada universalmente [59]. La Real Academia Española, en su 22° edición, define a la emoción como:

> f. Alteración del ánimo intensa y pasajera, agradable o penosa, que va acompañada de cierta conmoción somática[1].

Esta definición establece la idea que la emoción provoca cambios físicos en la persona que la siente. El diccionario de Oxford, por su parte, define al inglés *emotion* como:

> Un sentimiento fuerte que deriva de las propias circunstancias, el estado de ánimo o las relaciones con otros[2].

[1] `http://lema.rae.es/drae/?val=emocion` consultado el 17/03/2014.

[2] `http://www.oxforddictionaries.com/definition/english/emotion` consultado el 17/03/2014. Traducción propia.

En esta definición se hace hincapié en los orígenes de la emoción, y sólo se define a la emoción como un "sentimiento fuerte". La noción de "sentimiento" está muy relacionada a la de emoción, y es veces usada indistintamente. La Real Academia Española la define como:

> m. Estado afectivo del ánimo producido por causas que lo impresionan vivamente.

Estas definiciones, sin embargo, no nos permiten profundizar sobre los diversos aspectos que caracterizan a la emoción. Hay definiciones que enfatizan las características físicas, los mecanismos fisiológicos, definiciones que se centran en los estímulos que la producen, otras que la distinguen de otros procesos psicológicos, etc [59]. Kleinginna (1981) provee una definición que intenta abarcar varios de estos aspectos, y que consideramos apropiada a nuestros propósitos [59]:

> La emoción es un conjunto complejo de interacciones entre factores subjetivos y objetivos, mediados por sistemas hormonales y neuronales, que pueden (a) generar experiencias afectivas tales como sentimientos de excitación, placer o desagrado; (b) generar procesos cognitivos como efectos perceptuales emocionalmente relevantes, valorizaciones y procesos de calificación; (c) activar ajustes fisiológicos a las condiciones excitantes; y (d) conducir a comportamientos que usualmente son expresivos, dirigidos a las metas y adaptativos[3].

Chóliz [19] describe a la emoción desde diferentes perspectivas o teorías: evolucionistas, psicofisiológicas, neurológica, conductista, la teoría de la activación general y la cognitiva. La *psicofisiológica* arranca de la concepción de James (1884) y establece que la emoción aparece como consecuencia de la percepción de los cambios fisiológicos producidos por un determinado evento. En el caso de que no existan tales percepciones somáticas la consecuencia principal sería la ausencia de cualquier reacción afectiva. La teoría de las *estructuras neurológicas* es impulsada por Cannon (1931) en contraposición a la perspectiva psicofisiológica; se cuestiona el hecho de que dichas reacciones fisiológicas fueran un antecedente de la reacción emocional. Las reacciones fisiológicas y viscerales no definirían la cualidad de la reacción emocional, sino en todo caso la intensidad de la misma. Por lo tanto Cannon establece que lo verdaderamente relevante en la génesis de la emoción es la actividad del sistema nervioso central, en concreto la regulación que establece el tálamo, tanto sobre la corteza en la génesis de la experiencia cualitativa de la emoción, como sobre el sistema nervioso periférico, para la movilización de energía. En la teoría *conductista* (Watson, 1920) se entiende a la emociones como patrones de conducta relativamente fijos y no aprendidos, las manifestaciones externas de estos patrones de conducta hereditarios pueden romperse, o inhibirse parcialmente, a través del condicionamiento. No obstante, sus manifestaciones internas permanecen. Las emociones tienen un efecto disruptivo produciendo un cierto caos en la conducta. La *teoría de la activación general* (Lindsley, 1951)

[3]Traducción propia.

argumenta que existe un único estado de activación general que caracterizaría a todas las emociones. Las diferencias entre unas y otras sería cuestión de grado. La perspectiva *cognitiva* (Schachter, 1962) afirma que la emoción se produce por la conjunción de la activación corporal y de la interpretación cognitiva que la persona hace de esa activación corporal. La falta de uno de estos dos factores hace que la emoción sea incompleta. Schachter distingue entre dos tipos de experiencia emocional: una proveniente de las cogniciones del sujeto sobre la forma en que interpreta la situación que ha producido la emoción (esta experiencia se produce de manera rápida y bien diferenciada). El otro tipo de experiencia emocional proviene de las sensaciones corporales que produce la emoción, se trata de una experiencia lenta y bastante difusa.

En este trabajo se utiliza el enfoque evolucionista en donde se habla de emociones primarias de acuerdo a la "teoría de la paleta de colores" que establece que cada emoción puede ser descompuesta en emociones primarias de la misma forma en que ocurre con los colores. Las emociones primarias o universales son ira, asco, miedo, felicidad, tristeza y sorpresa [23, 29, 33]. Esta teoría deriva directamente de los planteamientos de Darwin (1872) y trata a las emociones como reacciones afectivas innatas, distintas entre ellas, presentes en todos los seres humanos, y que se expresan en forma unívoca [19, 29, 30, 33]. Los estudios en culturas occidentales y orientales, y en culturas no letradas, mostraron que estas emociones generan expresiones faciales que dependen de razones neurofisiológicas, y no de la cultura [33]. Para que una emoción sea considerada universal, debe cumplir con las siguientes condiciones [19, 53]:

- Tener un sustrato neural específico y distintivo.
- Tener una expresión o configuración facial específica y distintiva.
- Poseer sentimientos específicos y distintivos.
- Derivar de procesos biológicos evolutivos.
- Manifestar propiedades motivacionales y organizativas de funciones adaptativas.

Diversos estudios realizados por Schubiger [87] y O'Connor y Arnold [72] han establecido que existen 300 estados emocionales formados por las seis emociones básicas o universales antes mencionadas. Esta conceptualización de la emoción responde a la teoría evolucionista y sus principales pensadores son Charles Darwin, Robert Plutchik, Silvan Tomkins, Paul Ekman y Klauss Scherer.

1.2. Emociones en la voz y en las expresiones faciales

Definimos en la sección anterior qué se entiende por emoción o estado emocional, y en esta sección abordaremos el tema de cómo se relaciona la emoción con la voz y con las expresiones faciales. Las teorías evolucionistas indican que tanto la comunicación

vocal de la emoción como su comunicación a través de las expresiones faciales tienen un origen biológico y han evolucionado filogenéticamente[4] de manera continua [31, 82]. Estos indicadores de emoción dependen de factores biológicos (el tracto vocal, la vibración de las cuerdas vocales, los músculos faciales, etc.) y por lo tanto pueden ser estudiados objetivamente y traducibles en parámetros [1, 32].

Klauss Scherer y Paul Ekman han sido los principales o más reconocidos psicólogos que desde los años '70 han producido trabajos que han continuado y profundizado con las teorías darwinistas de la emoción y de su comunicación en la conducta vocal y las expresiones faciales, respectivamente. Es necesario distinguir entonces entre los trabajos psicológicos y sociales al respecto, y los trabajos relacionados a las ciencias de la computación, o a la inteligencia artificial. Los primeros han sentado las bases de los conocimientos actuales referidos a la percepción de la emoción: su universalidad, sus raíces biológicas y evolutivas, y los diferentes aspectos sociales y culturales que la condicionan [29, 31, 83, 84]. Desde el área de la computación se han tomado estos trabajos con el fin de resolver problemas prácticos y de interés para la sociedad.

A continuación se hará una introducción a las estructuras biológicas que participan en la producción de la voz, y se mencionarán algunas características de interés en esa señal. Nos centraremos en el aparato fonador, dejando de lado los procesos neuronales que intervienen en el proceso de producción de habla. En la Sección 3.1 se profundizará sobre los aspectos relacionados al modelado de la señal para su análisis computacional.

Para entender cómo se realiza la comunicación por medio de la voz, y en particular la comunicación de las emociones, no basta con analizar la producción de la voz. Es necesario analizar la forma en que el oído procesa esta señal, cómo se realiza la conversión de señales de audio en señales nerviosas y cómo procesa el cerebro esta información para hacer inteligibles a las ondas sonoras. Por lo tanto en la siguiente sección se hará un repaso sobre los conceptos básicos de la fisiología de la audición que son relevantes para el desarrollo de este trabajo.

Finalmente, en el último apartado de este capítulo, se mencionarán las relaciones entre las emociónes y las expresiones faciales y como éstas comunican fiablemente información con contenido emocional.

1.2.1. El aparato fonador y la señal de voz

El aparato fonador está constituido por los órganos que intervienen en la producción del habla, como se ven en la Figura 1.1. El tracto vocal está formado por la zona que está entre la laringe (glotis) y los labios, e incluye a las cavidades supraglóticas, faríngeas, oral y nasal. El aparato fonador se puede considerar un sistema que transforma energía mecánica (muscular) en energía acústica. En la teoría de señales y sistemas, muchas veces se lo simplifica considerándolo un sistema lineal invariante en el tiempo. La invariancia

[4]La filogenia se refiere al estudio de la historia evolutiva de especies o grupos de organismos. Para mayor información sobre este tema se puede consultar el libro de Wiley [105].

en el tiempo es por tramos, y se supone que está en los milisegundos que dura una cierta configuración del aparato fonador para producir un sonido determinado. Este sistema se modela como la respuesta de un sistema de filtros a una o más fuentes de sonidos.

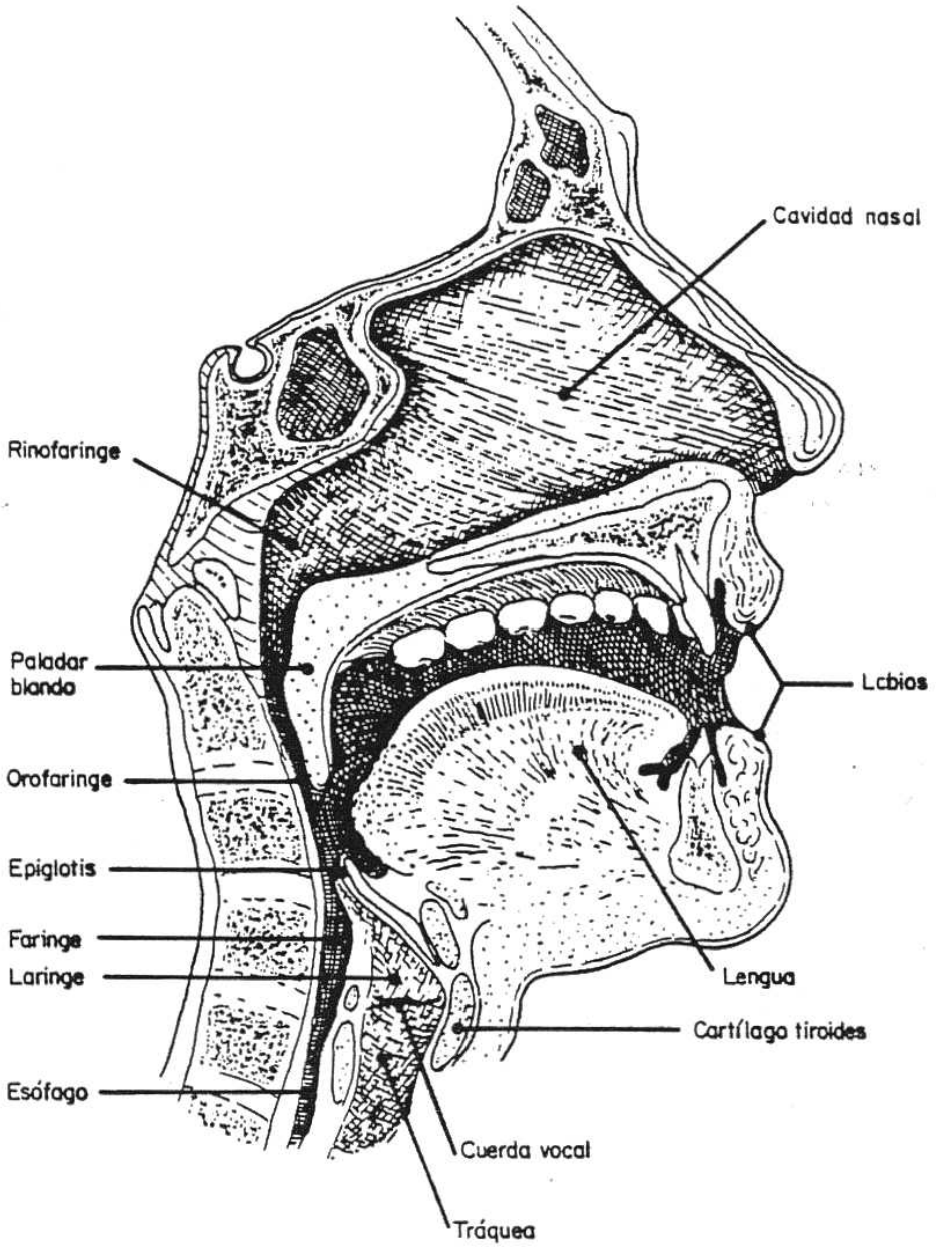

Figura 1.1: Corte sagital anatómico del aparato fonador (no incluye los pulmones) [5]

El sistema respiratorio proporciona energía en forma de flujos de aire y presiones que, a partir de las distintas perturbaciones del tracto vocal, genera los diferentes sonidos. Se pueden identificar tres mecanismos generales en la excitación del tracto vocal:

[5] Imagen obtenida de `http://liceu.uab.es/~joaquim/phonetics/fon_produccio/articulacion.html`

1. Las cuerdas vocales modulan el flujo de aire proveniente de los pulmones, lo que genera pulsos cuasi-periódicos.

2. Cuando este flujo de aire pasa por una constricción en el traco vocal se genera un ruido de banda ancha.

3. El flujo de aire produce una presión en un punto de oclusión total en el tracto vocal; cuando esta presión se libera rápidamente debido a la apertura de la constricción, se produce una excitación plosiva, intrínsecamente transitoria.

El aparato respiratorio es el responsable de la regulación de parámetros como la energía, la frecuencia fundamental de los pulsos cuasi-periódicos, el énfasis y la división del habla en unidades [81].

La laringe también interviene en la fonación, junto con las cuerdas vocales y los cartílagos donde se insertan. La vibración de las cuerdas vocales se modifica de forma voluntaria, y es la responsable de la frecuencia fundamental (F_0) del habla. Los valores de esta vibración varían entre 100 y 170 Hz en los hombres y entre 180 y 280 Hz en las mujeres, de ahí el sonido más agudo característico de la voz de la mujer.

El tracto vocal puede actuar sólo como modulador de los tonos glóticos si mantiene una configuración abierta, o estrechar o cerrar el paso de aire en una zona específica. Esto puede observarse en los espectros de los sonidos de las vocales, que proporcionan todos los aspectos relevantes de la configuración del tracto en ese instante [81]. Observando el espectro de cada vocal, se pueden distinguir las distintas formantes—picos de intensidad—con las cuales es posible diferenciarlas.

Las formantes corresponden a las resonancias del tracto vocal, y se numeran a partir del 1; F_1 y F_2 son principalmente un medio para caracterizar las vocales en español. La presencia de estas formantes y de la frecuencia fundamental (F_0) denota la presencia de un sonido sonoro o sordo (con o sin componente glótica) [81]. La frecuencia fundamental está directamente relacionada con la entonación de una frase.

La energía de corta duración es una medida fácilmente calculable que también es parte esencial de la entonación, y por lo tanto relevante al contenido emocional de una frase.

1.2.2. Fisiología de la audición

Nuestro sistema auditivo realiza con excelencia el proceso de decodificar las ondas sonoras para hacer inteligible la comunicación. A nuestro oído llegan las ondas sonoras provenientes de la voz de nuestro interlocutor, mezcladas con toda clase de ruidos, y nuestro sistema auditivo es capaz de funcionar correctamente frente a estas condiciones. Además, es capaz de entender los mensajes (y entender el contenido emocional de los mismos) independientemente de la pronunciación (más rápida o más lenta) y de la identidad del hablante. En la Figura 1.2 se puede observar un corte transversal del oído.

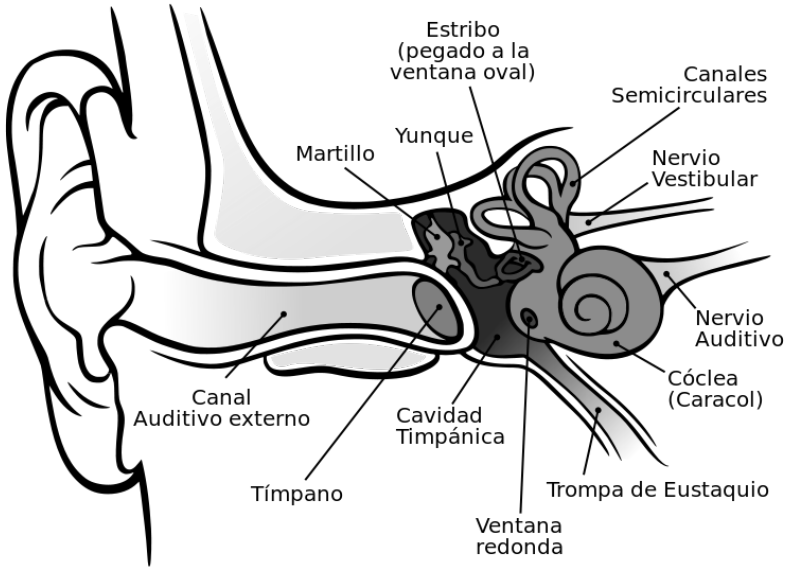

Figura 1.2: Corte sagital anatómico del oído.[6]

El oído está compuesto por tres secciones principales: el oído externo, el medio y el interno. El oído externo abarca desde el *pabellón auditivo* hasta el *tímpano*, y se encarga de captar el sonido y enfocarlo hacia el *conducto auditivo*, hasta llegar hasta el tímpano.

El tímpano es una membrana que vibra con las ondas de presión del aire, y transmite esta vibración al oído medio, que es una cámara aérea que contiene un conjunto de tres huesecillos: el *martillo*, el *yunque* y el *estribo*. Este último transmite la vibración a la *ventana oval* ubicada en la base de la *cóclea*, que constituye el oído interno [81].

La cóclea es similar a un tubo cónico lleno de un líquido, enrollado en forma de caracol. En su interior se encuentra la *membrana basilar*. En la Figura 1.3 se puede observar la ubicación de la cóclea en el oído interno

El funcionamiento del oído interno es muy complejo, pero lo más importante es que en él se realiza la transducción de la señal mecánica (de impulsos de aire, o vibraciones de huecesillos) en una señal eléctrica, que será conducida por el nervio auditivo al cerebro para ser procesada. Más precisamente, la transducción se realiza en la membrana basilar, a lo largo de la cóclea, y cuyas características son de vital importancia para entender cómo procesamos el sonido.

La membrana varía su rigidez a lo largo de su eje longitudinal, de mayor rigidez en la base, donde su ancho es mínimo, a menor rigidez en el ápex. Esto genera que las vibraciones de frecuencias más altas tienen mayor amplitud en el lugar donde las ondas comienzan a desplazarse, se atenúan en el camino y no alcanzan nunca el ápex. Por el contrario, las ondas de baja frecuencia tienen una pequeña amplitud en la base de la

[6]Imagen obtenida de `http://commons.wikimedia.org/wiki/File:Anatomia_del_Oido_humano.svg`

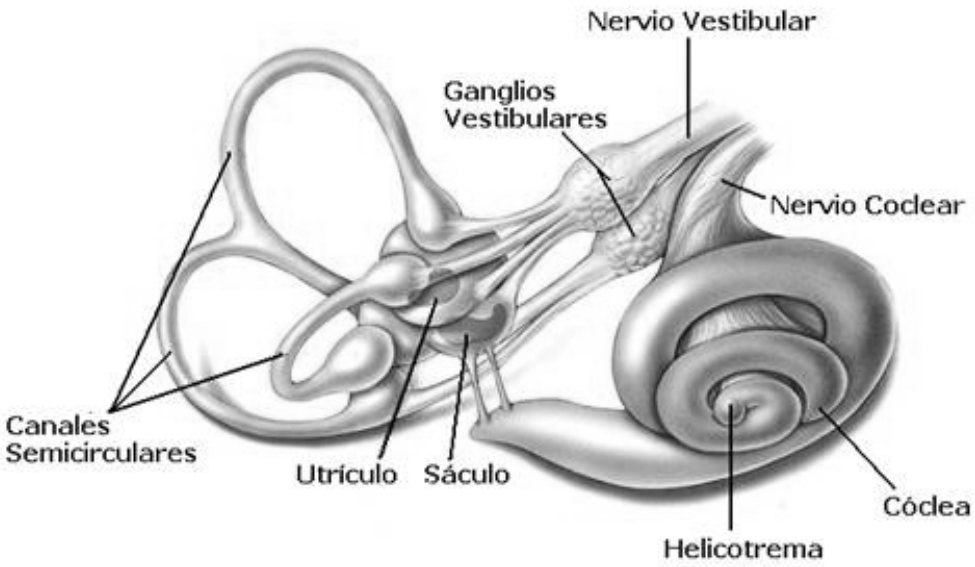

Figura 1.3: Ilustración del oído interno.[7]

membrana, y la aumentan mientras se acercan al ápex [81].

Debido a esta característica de la membrana y a otras causas, la resolución frecuencial y la percepción de las frecuencias no es uniforme a lo largo de la cóclea. La relación entre la distancia al estribo y la frecuencia de vibración máxima no es lineal, sino del tipo logarítmica. A esta escala psicoacústica que da cuenta de la relación entre la frecuencia física del sonido y la percibida se la denomina *escala de mel*. También se han realizado experimentos que dan cuenta de una escala similar (de carácter logarítmico) en la percepción de la intensidad de los sonidos [81].

Estas últimas características son de vital importancia en la representación de la señal por medios digitales para su análisis, como se detallará más adelante.

1.2.3. Las expresiones faciales como comunicadoras efectivas de la emoción

A partir de los años '60, principalmente Paul Ekman comenzó con investigaciones psicológicas en la línea de la teoría evolucionista. Antes de sus investigaciones, se creía que las expresiones faciales no presentaban ningún tipo de información confiable [30]. Se consideraba que las expresiones faciales eran como fonemas de un lenguaje: se creía que las unidades de comunicación estaban relacionadas a eventos específicos y experiencias de una manera específica como parte de la construcción cultural de la emoción. Décadas después, la comunidad científica ha aceptado que las expresiones faciales, a pesar de estar expuestas a modificaciones consecuencia de la cultura, no tienen un origen cultural. La

[7]Imagen obtenida de `http://en.wikipedia.org/wiki/File:Blausen_0329_EarAnatomy_InternalEar.png`

teoría evolucionista indica que las mismas son universales y que tienen un origen biológico [32, 33].

El hecho de que las expresiones faciales comuniquen fiablemente información acerca de la emoción ha sido fuente de debate a lo largo de las últimas décadas. Este debate se reduce a dos temas. Primero: si expresiones faciales inequívocas se corresponden a otros indicadores de emoción. Segundo: si observadores pueden juzgar las expresiones faciales de la emoción con precisión [32]. En la actualidad ya se han documentado relaciones consistentes y sustanciales entre las expresiones faciales y otros indicadores de emoción [32].

Se ha documentado que diferentes expresiones faciales voluntarias se relacionan con diferentes marcadores de actividad relevante a la emoción del Sistema Nervioso Autónomo (SNA) [62] y diferentes patrones de actividad del sistema nervioso central [34]. Por otro lado, expresiones faciales espontáneas se relacionan con diferentes respuestas del SNA en el caso de ira, compasión y risa [28, 56, 80]. La vergüenza, que también tiene su expresión facial particular, ha sido asociada con el rubor, que difiere con la respuesta autónoma de otras emociones [32].

De manera consistente con la teoría de que las expresiones faciales evolucionaron para provocar comportamientos específicos en una misma especie [32], la evidencia reciente muestra que efectivamente esto sucede. Las expresiones faciales evocan respuestas bastante específicas en los observadores [32]. Por ejemplo, la expresión facial de ira evoca respuestas autónomas relacionadas al miedo que son distintas a las respuestas provocadas por sonrisas [35]. Se ha mostrado que las expresiones faciales de angustia provocan compasión [28], y que las expresiones de ruborización y vergüenza provocan emociones de diversión y compasión respectivamente [57]. Las expresiones faciales de diferentes emociones negativas evocan diferentes emociones en observadores, lo que concuerda con el acercamiento discreto a la descripción de las emociones [32]. Más aún, estudios recientes muestran que los patrones de actividad cerebrales como respuesta a los emoticones usados en la comunicación virtual son similares a los patrones de activación observados ante expresiones faciales reales [20, 109].

Expresiones faciales universales asociadas a sus respectivas emociones

En la comunidad científica hay dos tendencias para modelar las emociones. La primera consiste en un modelo discreto de la emoción, en el cual hay emociones como alegría, tristeza, sorpresa, ira, etc. La segunda, que ya anticipamos al principio de este capítulo, trata las emociones en un modelo bidimensional de valencia y activación. En la última década, se ha puesto mayor interés en el segundo modelo a medida que se ha acrecentado el estudio en el reconocimiento automático de emociones "espontáneas", ya que permite mayor flexibilidad al mostar los cambios continuos en la presencia o ausencia de emoción [111].

Sin embargo, no hay que olvidar que ambos son modelos y que cada uno puede ser más o menos adecuado para la aplicación que sea necesaria. En el trabajo de Ekman,

"Expresiones faciales de la emoción" (1993), se muestra evidencia de que las expresiones faciales son percibidas categóricamente y están relacionadas a regiones cerebrales distintas, actividad del SNA, y respuestas evocadas en terceros [32]. Es decir, nosotros sentimos y percibimos las emociones, al menos relacionadas con las expresiones faciales, de manera discreta y categórica. Sin embargo, a pesar de que podemos distinguir emociones discretas, también es cierto que percibimos las emociones con diferentes intensidades (por ejemplo: enojado, muy enojado, etc.). Es por esto, que el modelo de activación y valencia resulta útil en los casos en los que se quiere analizar la emoción sentida instantáneamente [32].

Ekman realizó estudios en diferentes culturas letradas y no letradas, occidentales y no occidentales, y concluyó que las emociones universales que tienen sus expresiones faciales características y únicas son seis: ira, asco, miedo, felicidad, tristeza y sorpresa [33]. En la Figura 1.4 se pueden ver las expresiones faciales de estas emociones.

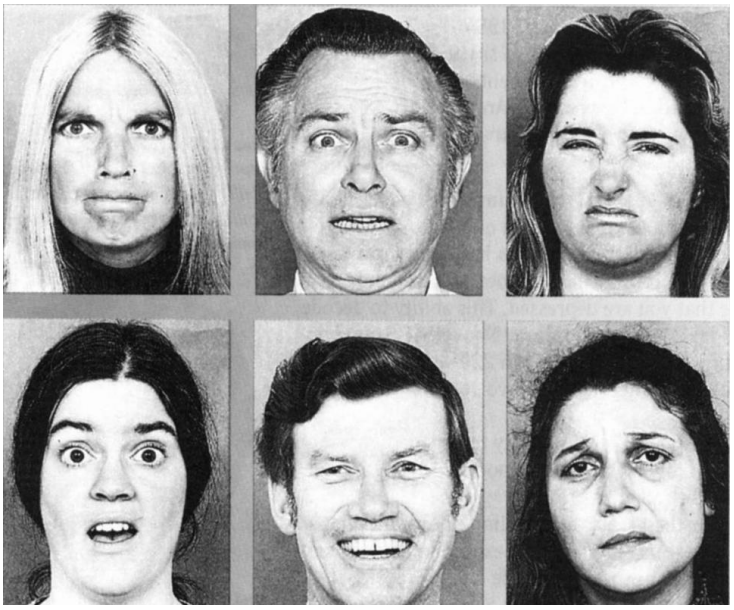

Figura 1.4: Imágenes de las expresiones faciales. De izquierda a derecha y de arriba a abajo: ira, miedo, asco, sorpresa, alegría y tristeza.[8]

Para explicar las diferencias culturales percibidas en las expresiones faciales, Ekman propuso una teoría, que llamó *neurocultural*, la cual es aceptada en la comunidad científica. Según esta teoría, las diferencias culturales ocurren: (a) porque muchos de los eventos que inducen las emociones son aprendidos y varían culturalmente, (b) porque las reglas para controlar las expresiones faciales en determinadas situaciones sociales varían culturalmente, y (c) porque algunas consecuencias de la excitación emocional también varían en las distintas culturas [33].

[8]Imagen obtenida de `http://www.bostonglobe.com/ideas/2012/02/11/emoticon-your-face/NoAXTpkjCPSHSzpdJjqeOJ/story.html`

1.3. Justificación

En los últimos años se ha puesto mucho esfuerzo en la comprensión automática del discurso [64, 96], en la detección de emociones en el discurso [3, 7] y en la detección de emociones en expresiones faciales [38]. Se han propuesto diversos métodos que usan diferentes modelos para intentar solucionar estos problemas. Estos acercamientos monomodales no analizan la correlación entre los distintos canales de comunicación. Se desprende aquí la necesidad de explotar la correlación entre los canales más relacionados con el transporte de información emocional en la comunicación humana: la voz y las expresiones faciales. El análisis conjunto de estos canales permitiría hacer un estudio multimodal de la comunicación emocional y mejorar la capacidad de los sistemas convencionales de discernir entre emociones. Es por ésto que la tendencia actual es analizar conjuntamente diferentes modalidades, como lo son la señal de voz, los gestos, las expresiones faciales, etc. [14, 41, 112].

1.4. Objetivos

1.4.1. Objetivo general

- Diseñar un sistema multimodal de identificación de emociones a partir del discurso y las expresiones faciales.

1.4.2. Objetivos específicos

- Relevar el estado del arte de las soluciones actuales a problemas similares.
- Estudiar las bases de datos multimodales (audio/video) de emociones existentes a la fecha y seleccionar la más apropiada para el desarrollo y prueba del sistema.
- Desarrollar un método para la extracción de características en expresiones faciales registradas en video.
- Desarrollar un método para la extracción de características en la señal de voz.
- Diseñar un clasificador para lograr el reconocimiento multimodal de emociones.
- Evaluar el desempeño de los métodos desarrollados, comparar con métodos del estado del arte y analizar resultados.

Capítulo 2

Estado del arte

En este capítulo se hará una revisión del estado del arte en el reconocimiento automático de emociones en señales de video de expresiones faciales, en señales de voz y en señales multimodales con ambas fuentes de información. Este análisis se realiza con el objetivo de conocer el estado del conocimiento científico en la materia, evaluar las alternativas propuestas en la comunidad científica para la solución de problemas similares y luego proponer una solución propia.

El problema de reconocimiento automático de emociones–tanto en expresiones faciales, como en voz o señales multimodales–es un problema de reconocimiento de patrones, que es un tipo de problema característico dentro de las disciplinas de Aprendizaje de Máquina o de Inteligencia Computacional. Este problema consiste en asignar una etiqueta a un valor de entrada, cuya etiqueta (tipo) se desconoce. En este caso, el valor de entrada, o patrón, sería una expresión facial, o una señal de voz, o una señal multimodal, y la etiqueta es la emoción presente en esa señal.

Este tipo de problema consiste en dos etapas generales. La primera, llamada extracción de características, responde qué subconjunto de datos de cada patrón es relevante para que el sistema pueda obtener información, o qué tipo de transformación sobre los datos de entrada hay que realizar para obtener información significativa de estos patrones, y que permita discriminar las distintas clases. El objetivo de esta etapa es encontrar una forma de representación de los datos de entrada que maximice las diferencias entre patrones de diferentes clases, y minimice las diferencias entre patrones de la misma clase.

La segunda etapa es la clasificación, y consiste en determinar a qué clase pertenece un patrón dadas las características obtenidas en la etapa anterior. Un clasificador puede verse como un modelo matemático; el cual ofrece una salida o etiqueta en función de una determinada entrada. Hay distintos tipos de clasificadores: estadísticos, basados en redes neuronales o en distintos modelos matemáticos, entre otros. Además, hay distintas formas de codificar la información en los clasificadores, y de hacer que que el clasificador "aprenda" esta información.

2.1. Reconocimiento de emociones en expresiones faciales

A partir de la década de los '90, con el crecimiento del interés en Interacción Humano-Computadora, del inglés *Human–Computer Interaction* (HCI) y *Affective Computing*[1], numerosos grupos de investigación comenzaron con el estudio del reconocimiento automático de emociones en expresiones faciales.

En esta sección se hará una revisión de los trabajos más actuales, realizados principalmente a partir de la década del '00. Trabajos realizados con anterioridad se pueden revisar en las recopilaciones de Fassel et al. [38] y Pantic et al. [76].

En la etapa de la extracción de características, los métodos más utilizados son los siguientes: MPEG-4 Parámetros de Animación Facial, del inglés *Facial Animation Parameters* (FAPs) extraídos utilizando el algoritmo Active Contour y estimación de movimiento [77], *Gabor Wavelets* [9, 61], vector de desplazamiento de características (distancia euclídea entre emoción neutral y pico de emoción) [67], rastreo de un conjunto de 20 puntos faciales [75], 34 puntos faciales convertidos en Grafos Etiquetados, del inglés *Labelled Graph* (LG) usando la transformada onditas de Gabor, luego construcción de un vector semántico de expresión para cada cara de entrenamiento y Análisis de Correlación de Núcleo Canónico, del inglés *Kernel Canonical Correlation Analysis* (KCCA) usado para aprender la correlación entre el vector LG y el vector semántico [119], firmas de movimiento obtenidas por rastreo usando *spatial ratio template tracker* y realizando flujo óptico en la cara usando un gradiente multicanal (MCGM en inglés) [6], MPEG-4 FAPs del contorno de la boca (grupo 8) y de cejas (grupo 4) seguido de Análisis de Componentes Principales, del inglés *Principal Component Analysis* (PCA) para reducir dimensionalidad [5].

Los MPEG-4 FAPs son un conjunto de 86 parámetros definidos en el estándar ISO/IEC ISO/IEC 14496-1 y 2 por el Consorcio MPEG[2] para la representación virtual de humanos y humanoides en una forma que proporcione inteligibilidad visual del discurso y de las expresiones faciales y gestuales del hablante.

La Transformada de onditas Gabor (*Gabor wavelets*) es frecuentemente utilizada en la extracción de características de imágenes, debido a su capacidad de compresión y representación de las imágenes en forma rala (pocos coeficientes).

PCA es una técnica estadística muy utilizada para reducir la dimensionalidad de las características posiblemente correlacionadas. La transformación está diseñada de una manera en que la que la primer componente maximice la varianza, y las siguientes tengan la mayor varianza posible. De esta manera, se pueden representar los datos en menor cantidad de dimensiones proyectándolos sobre este nuevo espacio.

[1] Se denomina Computación Afectiva o *Affective Computing* al estudio y desarrollo de sistemas de reconocimiento, interpretación, procesamiento y simulación computacional de emociones.

[2] `http://mpeg.chiariglione.org/standards/mpeg-4/video`

Para la clasificación, son utilizados los siguientes métodos: Modelos Ocultos de Markov, del inglés *Hidden Markov Models* (HMM) [5, 77], Modelos Ocultos de Markov Multietapa, del inglés *Multi Stage Hidden Markov Models* (MS-HMM) [5], Máquinas de Soporte Vectorial, del inglés *Support Vector Machines* (SVM) [6, 9, 67], *AdaBoost SVM* (AdaSVM) [9], SVM multiclase [61], reglas temporales [75], vector semántico de expresiones [119] y Perceptrón Multicapa, del inglés *Multilayer Perceptron* (MLP) [6, 61].

Los modelos ocultos de Márkov (HMM por sus siglas en inglés) son modelos estadísticos en los que se supone que el proceso a modelar es un proceso de Márkov cuyos parámetros son desconocidos, y es preciso estimar. Un proceso de Márkov es un fenómeno aleatorio dependiente del tiempo que cumple ciertas características. Estos parámetros son los que son utilizados para el reconocimiento de patrones. Los modelos ocultos de Márkov son utilizados para aprovechar la temporalidad en la expresión de las emociones.

En el Capitulo 3 se profundizará sobre los métodos de clasificación utilizados en este trabajo.

Los trabajos revisados en el análisis del estado del arte utilizan las siguientes bases de datos para diseñar y probar sus modelos: Cohn-Kanade[3] [5, 9, 61, 67, 75, 77], JAFFE[4] [61, 119], MMI[5] [75], Imágenes Afectivas de Ekman [119] y CMU-Pittsburg codificada con AU [6].

2.2. Reconocimiento de emociones en la voz

Se pretende presentar los métodos más usados de extracción de características de señales de voz. Con este objetivo, es necesario hacer una separación de las características en clases, aunque no hay una sóla manera de hacer esto. La primer gran división es de origen tecnológico: las características acústicas y las características lingüísticas se separan ya que los métodos de extracción para estos dos tipos son muy diferentes [88]. Su contribución de información varía según la base de datos y la aplicación. En el reconocimiento automático de emociones se utilizan las características acústicas, y en la comprensión automática del discurso se utilizan las características lingüísticas [88].

Las características acústicas más utilizadas en son las características prosódicas tales como el *pitch*, la F_0, la duración y la intensidad (energía) [4, 42, 52, 55, 70, 84, 106]; las características de calidad de la voz, tales como el Relación Armónicas-Ruido, del inglés *Harmonics-to-Noise Ratio* (HNR), el *jitter* y el *shimmer* [63, 84, 106]; las características espectrales [4], formantes de modelado y características cepstrales como los MFCC [4].

El *pitch* es una medida perceptual de la frecuencia fundamental F_0. Para el cálculo de la energía se utiliza la energía de corta duración (por ventanas) y luego se analiza su variabilidad o su promedio.

[3] `http://www.pitt.edu/~emotion/ck-spread.htm`
[4] `http://www.kasrl.org/jaffe.html`
[5] `http://www.mmifacedb.com/`

El HNR es una medida de la relación señal-ruido de señales periódicas o cuasi-periódicas. Es una medida muy utilizada en estudios sobre calidad de voz, envejecimiento, ronqueras, enfermedades, etc. [40, 90, 110]. El *jitter* se refiere a la variabilidad en el *pitch* de la voz, período a período, lo que causa un sonido áspero. El *shimmer* se refiere a los cambios de intensidad en la voz período a período.

Las formantes de modelado son las que describimos en la Sección 1.2.1. El *cepstrum* de una señal es el resultado de calcular la Transformada de Fourier del espectro de la señal en escala logarítmica. En la Sección 3.1 se profundizará sobre los MFCC.

En cuanto a la clasificación, el estado del arte es usar HMM si se opta por clasificadores dinámicos. También se han usado SVM, ANN (en general MLP), métodos de ensamble [69, 89] y clasificación jerárquica [4].

2.3. Reconocimiento multimodal de emociones

Para realizar el reconocimiento de emociones en señales multimodales, es necesario extraer tanto características de la señal de audio como de la de video.

Luego, se han realizado diferentes acercamientos a la fusión de la información. Existen tres alternativas: fusión de los datos a nivel de características [14, 113], fusión de los datos a nivel de decisión [14, 44, 50, 101, 112, 115, 116] y fusión a nivel de modelo [41, 114, 117].

Estudios neurológicos en la fusión de neuronas sensoriales apoyan una fusión de datos temprana; es decir, fusión a nivel de características [94]. Sin embargo, cómo fusionar estos datos es algo que todavía está en discusión [111].

Debido a esta dificultad, la mayoría de los investigadores han optado por fusión a nivel de decisión, en la cual cada modo de señal se modela independientemente, y los resultados de la extracción de características y clasificación de estas señales independientes se fusionan al final. En este tipo de fusión, la clasificación no aprovecha la correlación de los datos, dado que se realiza antes de la fusión de los mismos y de manera independiente.

La fusión a nivel de modelo es un acercamiento híbrido, que intenta aprovechar la correlación de las señales y la sencillez de implementación de una fusión tardía.

Los métodos más utilizados para la extracción de características son: marcadores en la cara [14], FAPs [41], Análisis de Discriminante Lineal, del inglés *Linear Discriminant Analysis* (LDA) [44], PCA [98], *Gabor wavelets* [50, 101], unidades de movimiento [112-114, 116, 117], textura con LLP [115], características prosódicas [14, 41, 50, 101, 112-117], MFCC [44, 101] y formantes [101, 112].

La clasificación se realiza mediante los siguientes métodos: SVM [14, 50], Redes Neuronales Artificiales con bucle de retroalimentación, del inglés *Artificial Neural Network with a feedback loop* (ANNA) [41], creación de un *codebook* [44], Análisis de Discriminante Lineal de Fisher, del inglés *Fisher's Linear Discriminant Analysis* (FLDA) [101], HMM [112], Modelos Ocultos de Markov Multiflujo Fusionados, del inglés Multi-stream Fused

Hidden Markov Models (MFHMM) [114, 117], AdaBoost y Modelos Ocultos de Markov de Multiresolución, del inglés *Multiresolution Hidden Markov Models* (MHMM) [115], Red Rala de Winnows, del inglés *Sparse Network of Winnows* (SNoW) [116] y Fisher-Boosting [113].

LDA y FLDA son técnicas utilizadas para encontrar una combinación lineal de características que caracterice o separe dos o más clases de objetos o eventos [54].

La arquitectura de aprendizaje SNoW es un clasificador multiclase que está diseñado para tareas de aprendizaje de gran escala, en las que el número de características de entradas es muy elevado, y puede ser desconocido a priori [15].

En el Capitulo 3 se profundizarán los métodos de clasificación utilizados en este trabajo.

Capítulo 3

Fundamentos teóricos

En este capítulo se estudian los fundamentos teóricos en los que este trabajo se sustenta.

En la primera sección se repasan los conceptos básicos de señales y sistemas, y se dan fundamentos de los diferentes métodos para el análisis y procesamiento digital de la voz.

En la segunda sección se hace una introducción al análisis de componentes principales, técnica estadística de reducción de dimensionalidad que utilizaremos en la extracción de características de video.

Por último, en la tercer sección se estudian los conceptos referidos a las redes neuronales. Se repasa la teoría detrás del perceptrón simple y el perceptrón multicapa. Luego se detalla el concepto de aprendizaje, y se repasa el algoritmo de entrenamiento más utilizado, el algoritmo de retropropagación.

3.1. Procesamiento digital de la señal de voz

Comenzamos esta sección introduciendo los conceptos básicos de señales y sistemas. Luego, se repasan conceptos importantes para el procesamento de la voz, como son la Transformada Discreta de Fourier (TDF), los sitemas lineales e invariantes en el tiempo y el análisis cepstral de la señal.

3.1.1. Conceptos básicos de señales y sistemas

Los conceptos de *señales*, *sistemas* e *información* se enmarcan en la teoría de las comunicaciones. Se puede decir que las señales transportan información acerca del sistema que las produjo, contenida o codificada en un patrón de variaciones de alguna magnitud física. Esta magnitud física puede ser de cualquier clase: eléctrica, lumínica, sonora, magnética, de calor, etc. Como vimos anteriormente, el aparato fonador humano produce

los patrones de variación de la presión del aire que constituyen la señal sonora de la voz, la cual es la base de la comunicación humana [74, 81].

Las señales son descriptas matemáticamente por funciones y los sistemas por transformaciones. Las transformaciones modifican señales de entrada produciendo señales de salida.

Según su evolución en el tiempo las señales pueden ser *determinísticas*, si conocemos sus valores de antemano o podemos predecirlos exactamente, o *aleatorias*, si existe algún tipo de incerteza sobre los valores que puede tomar la señal.

Las señales físicas son por lo general continuas, y para que puedan ser tratadas por computadora es necesario realizar una *conversión analógica a digital*, que consiste en muestrear la señal y cuantizarla en amplitud. Esta conversión puede introducir dos tipos de *ruido*. El primer tipo es el debido a la cuantización en niveles discretos de la amplitud de la señal. Esto introduce imprecisión, ya que no se pueden tener infinita cantidad de bits para representar la magnitud de la señal. El segundo tipo se debe a la velocidad o *frecuencia de muestreo*. Si se muestrea la señal a una velocidad más lenta que el doble de la de mayor frecuencia presente en la señal se modifica la información existente el el rango útil. Este fenómeno se denomina *aliasing*, y para evitarlo es necesario aplicar el teorema del muestreo de Nyquist, que consiste en muestrear la señal al doble de frecuencia de muestreo que la mayor frecuencia presente en la señal.

Otro concepto importante es el de ruido en una señal. Se denomina ruido a cualquier fenómeno o proceso que altera la percepción o interpretación correcta de la señal. Una medida muy utilizada de cuánto una señal está contaminada por ruido es la Relación Señal Ruido (SNR), que es la razón entre la potencia de la señal P_s y la potencia del ruido P_r [81]:

$$\xi = \frac{P_s}{P_r}$$

Dado que las computadoras sólo pueden tratar con señales discretas, y estas señales discretas son, al fin y al cabo, una sucesión de números, se puede establecer una relación entre el álgrebra lineal y la teoría de señales. Las señales entonces pueden representarse como vectores, y toda la teoría del álgebra lineal puede aplicarse para su tratamiento. Decimos entonces que una señal en un espacio N-dimensional es un vector $[x_1, x_2, \ldots, x_N]$ definido como una N-tupla ordenada de números. Para estos vectores se utiliza la notación:

$$\mathrm{x} = [x_n]; n \in \mathbb{N}; x_n \in \mathbb{R}; \mathrm{x} \in \mathbb{R}^N$$

Estos conceptos son importantes porque podemos aprovechar los conceptos de ortogonalidad, combinación lineal e independencia lineal del álgebra en el tratamiento de señales.

Dos conceptos importantes en el análisis de señales, y que serán utilizados en el desarrollo de este trabajo, son la energía de la señal (E) y la frecuencia fundamental de la señal (F_0). La energía de la señal discreta x se define como:

$$E = \sum_{n=0}^{N} x[n]^2$$

La frecuencia fundamental es aplicable a señales periódicas, y se define como la frecuencia más baja de su espectro de frecuencias.

3.1.2. Transformada discreta de Fourier

La Transformada Discreta de Fourier (TDF) es una transformación muy importante y muy utilizada en la teoría de señales, y utiliza la base de exponenciales complejas generadoras de $\mathbb{R}^N$. Esta base está dada por:

$$\phi[k](t) = e^{j\frac{2\pi kn}{N}}$$

La TDF entonces se define como:

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j\frac{2\pi kn}{N}} \tag{3.1}$$

Esta transformación se aplica a señales discretas periódicas de duración infinita, o a extensiones periódicas de señales de duración finita, de manera que si hacemos esta transformación a la señal de voz estamos haciendo esta suposición. La señal de voz es una señal aleatoria cuasi-periódica cuando se producen sonidos sonoros (con componente glótica), por lo que puede aplicarse esta transformación teniendo esto en cuenta [74, 81].

Al realizar esta transformación, se dice que se está *descomponiendo* la señal en sus componentes frecuenciales, ya que se la está multiplicando mediante producto punto con exponenciales complejas de diferentes frecuencias. Esta multiplicación mide el *grado de parecido* de la señal con las exponenciales de distintas frecuencias. Al resultado de la TDF se lo conoce como *espectro* de la señal [74, 81].

3.1.3. Sistemas lineales e invariantes en el tiempo

Se puede definir a un sistema como cualquier proceso que realiza una transformación en la señal. Los sistemas lineales e invariantes en el tiempo (LTI) tienen propiedades que son aprovechadas en el tratamiento de las señales que producen. Se define este tipo de sistemas porque se asume que el aparato fonador se comporta como un sistema lineal e invariante en el tiempo durante los milisegundos que se mantiene una configuración determinada para producir un sonido. Hay que notar que ésta es una aproximación, ya que el aparato fonador real tiene componentes no lineales y además es claramente variante en el tiempo.

Un sistema es lineal si posee la propiedad de superposición. Es decir, si una entrada consiste en una suma ponderada de varias señales entonces la salida es la superposición (la suma ponderada) de las respuestas del sistema para esas señales. Esto permite descomponer una señal en señales más sencillas (exponenciales de Fourier por ejemplo) para el análisis de la respuesta del sistema.

Un sistema es invariante en el tiempo si un desplazamiento en la entrada produce el mismo desplazamiento en la salida.

En el contexto del análisis de señales, el sistema de producción de habla descripto en la Sección 1.2.1 se modela como un sistema lineal. Esto significa que la salida del sistema $y(t)$ puede obtenerse como convolución de la entrada $x(t)$ y la respuesta al impulso $x(t)$:

$$y(t) = x(t) * h(t) \tag{3.2}$$

En el análisis de la voz, se tiene acceso sólo a la señal $y(t)$, y frecuentemente es deseable eliminar alguna de las componentes $x(t)$ o $h(t)$ de manera de poder analizar la restante. Sin embargo, la operación de *deconvolución* es un problema inverso complejo que muchas veces es difícil de resolver, debido a la presencia de ruido.

Un método que permite resolver este problema de separar las señales combinadas mediante la convolución es el *análisis cepstral*.

3.1.4. El análisis cepstral

El *cepstrum* o cepstro de una señal es el resultado de calcular la TDF al espectro de la señal en escala logarítmica. El nombre *cepstrum* proviene de invertir las cuatro primeras letras de *spectrum*. Matemáticamente, se define al *cepstrum* $C_y(t)$ de la señal $y(t)$ como:

$$C_y(t) = \mathcal{F}^{-1}\{\log(\mathcal{F}\{y(t)\})\} \tag{3.3}$$

donde $\mathcal{F}\{\cdot\}$ es el operador de la TDF, y se supone que $y(t)$ es una señal generada por un sistema LTI [81]. Este método aprovecha las propiedades de la TDF y de logaritmo. La TDF de $y(t)$ es:

$$Y(f) = X(f)H(f)$$

con lo que se transformó una operación de convolución en un producto. Luego se aplica logaritmo para transformar un producto en suma, y se vuelve al dominio "temporal" (denominado dominio de las *cuefrencias*), aplicando la TDF inversa. La importancia del *cepstrum* es que la señal es ahora una combinación lineal de sus componentes, y que las mismas, si presentan diferentes características frecuenciales, estarán ubicadas en distintas *cuefrencias*, lo que nos permitirá separarlas.

El *cepstrum* definido como en la Ecuación 3.3 es denominado comúnmente Cepstrum Complejo (CC), mientras que el Cepstrum Real (CR) es más utilizado para el análisis de habla debido a su mayor simplicidad de uso. La diferencia es que el primero mantiene la información de la fase de las señales, mientras que el último la descarta [81].

Como se explicó en la Sección 1.2.1, la discriminación de frecuencias en nuestro oído no es lineal. Debido a esto, en el procesamiento de voz se utilizan bancos de filtros para las denominadas *bandas críticas* [81]. El tipo de filtro más utilizado es el de ventana triangular en la escala psicoacústica de mel. La relación entre la escala lineal en Hz. y la escala de mel se muestra en la Figura 3.1.

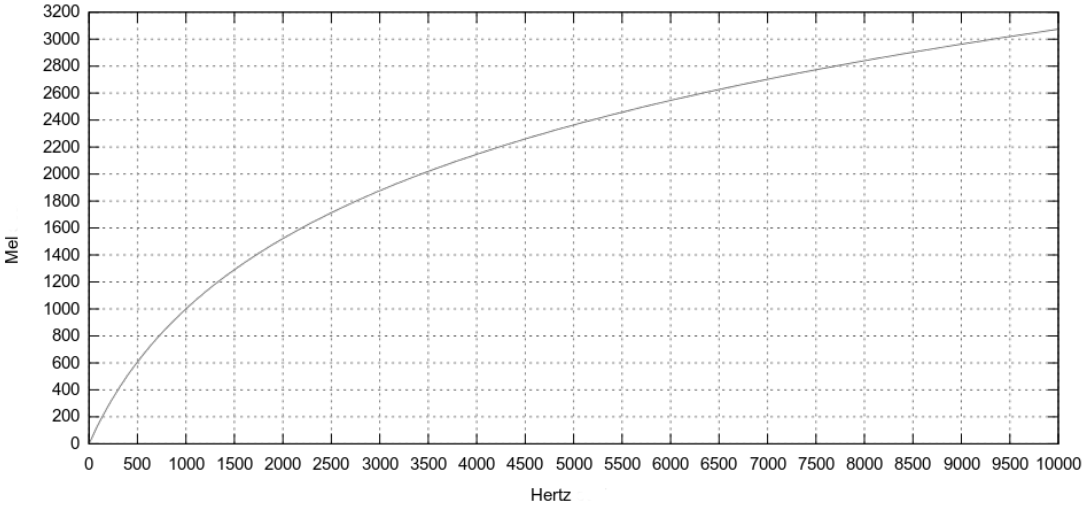

Figura 3.1: Escala de Mel.[1]

Esto muestra que la relación es aproximadamente lineal por debajo de 1 KHz y logarítmica por encima, lo que lleva a una aproximación muy utilizada [25]:

$$f_{mel} = \frac{1000}{\log(2)} \log\left[1 + \frac{f_{Hz}}{1000}\right]$$

Las técnicas trabajan utilizando el espectro de potencia y los coeficientes *cepstrales* de la señal, representación que se denomina *coeficientes cepstrales en escala de mel* (MFCC). Dado que tanto la amplitud y la forma del *cepstrum* son suceptibles al ruido y se modifican con un simple cambio de micrófono, para agregar robustez se suele utilizar el *delta cepstrum* (ΔC) [25]. El ΔC calcula la diferencia *cepstral* entre el segmento de voz actual y el anterior, y constituye una aproximación a la derivada temporal del *cepstrum*. Muchos sistemas utilizan el *cepstrum*, el ΔC e incluso la segunda derivada, el $\Delta\Delta$C [81].

3.2. Análisis de Componentes Principales

El Análisis de Componentes Principales, del inglés *Principal Component Analysis* (PCA) es una técnica estadística ampliamente utilizada en el área de reconocimiento de patrones y procesamiento digital de señales para lograr reducir dimensionalidad en los datos [47].

Un problema típico es el de seleccionar o extraer características relevantes en los datos. Más precisamente, la *extracción de características* refiere al proceso mediante el cual el espacio de los datos originales es transformado al espacio de características. El objeto aquí es diseñar una transformación que permita representar al conjunto de datos de origen por un reducido conjunto de características "relevantes" que contengan la información intrínseca de mayor importancia.

[1] Imagen obtenida de `http://en.wikipedia.org/wiki/File:Mel-Hz_plot.svg`

Lo que se busca entonces es encontrar una transformación lineal $\mathbf{T}$ que al aplicársela a un vector $\mathbf{x}$ optimice el error cuadrático medio que el truncamiento en los datos genera. PCA maximiza la tasa de decrecimiento de la varianza y por esto es la transformación lineal que buscamos.

Sea $\mathcal{X}$ un *vector aleatorio* de dimensión m que representa el entorno de interés. Asumimos que $\mathcal{X}$ tiene media cero. Esto es:

$$E[\mathcal{X}] = \mathbf{0}$$

donde E es el operador estadístico esperanza. Sea $\mathbf{q}$ un vector unitario también de dimensión m, en el cual el vector $\mathcal{X}$ será proyectado. Esta proyección está definida por el producto interno de ambos vectores, entonces:

$$A = \mathcal{X}^T\mathbf{q} = \mathbf{q}^T\mathcal{X} \tag{3.4}$$

sujeto a la condición:

$$||\mathbf{q}|| = (\mathbf{q}^T\mathbf{q})^{1/2} = 1 \tag{3.5}$$

La proyección A es una *variable aleatoria* con media y varianza relacionada con las estadísticas del vector $\mathcal{X}$. Como antes mencionamos, asumiendo que el vector $\mathcal{X}$ tiene media cero, la media de la proyección A es cero también,

$$E[A] = \mathbf{q}^T E[\mathcal{X}] = 0$$

La varianza de A es entonces igual a su valor cuadrático medio. Por lo tanto:

$$\begin{aligned} \sigma^2 &= E[A^2] \\ &= E[(\mathbf{q}^T\mathcal{X})(\mathcal{X}^T\mathbf{q})] \\ &= \mathbf{q}^T E[\mathcal{X}\mathcal{X}^T]\mathbf{q} \\ &= \mathbf{q}^T\mathbf{R}\mathbf{q} \end{aligned} \tag{3.6}$$

La matriz de $m \times m$ $\mathbf{R}$ es la matriz de *correlación* del vector aleatorio $\mathbf{X}$, definida formalmente como la esperanza del producto externo de X consigo mismo. Esto es:

$$\mathbf{R} = E[\mathcal{X}\mathcal{X}^T] \tag{3.7}$$

Se observa que la matriz de correlación $\mathbf{R}$ es *simétrica*, lo que significa que:

$$\mathbf{R}^T = \mathbf{R} \tag{3.8}$$

De esta propiedad se observa que si $\mathbf{a}$ y $\mathbf{b}$ son vectores $m \times 1$, entonces:

$$\mathbf{a}^T\mathbf{R}\mathbf{b} = \mathbf{b}^T\mathbf{R}\mathbf{a} \tag{3.9}$$

De la Ecuación 3.6 se puede notar que la varianza σ^2 de la proyección A es función del vector unitario $\mathbf{q}$. Por lo que podemos escribir:

$$\begin{aligned} \psi(\mathbf{q}) &= \sigma^2 \\ &= \mathbf{q}^T \mathbf{R} \mathbf{q} \end{aligned} \tag{3.10}$$

Se toma a $\psi(\mathbf{q})$ como un estimador de la varianza.

3.2.1. Autovalores y Autovectores en PCA

Lo que se busca ahora es encontrar esos vectores unitarios $\mathbf{q}$ en los cuales $\psi(\mathbf{q})$ tiene valores *extremos o estacionarios* (máximos o mínimos locales), sujetos a la restricción de la norma euclídea de $\mathbf{q}$. La solución de este problema se encuentra en el conjunto de autovalores de la matriz de correlación $\mathbf{R}$. Si $\mathbf{q}$ es un vector unitario tal que el estimador de la varianza $\psi(\mathbf{q})$ tiene un valor extremo, entonces para cualquier pequeña perturbación $\delta\mathbf{q}$ del vector $\mathbf{q}$, encontramos que:

$$\psi(\mathbf{q} + \delta\mathbf{q}) = \psi(\mathbf{q}) \tag{3.11}$$

Ahora, de la definición de la varianza dada en Ecuación 3.10, tenemos:

$$\begin{aligned} \psi(\mathbf{q} + \delta\mathbf{q}) &= (\mathbf{q} + \delta\mathbf{q})^T \mathbf{R} (\mathbf{q} + \delta\mathbf{q}) \\ &= \mathbf{q}^T \mathbf{R} \mathbf{q} + 2(\delta\mathbf{q})^T \mathbf{R} \mathbf{q} + (\delta\mathbf{q})^T \mathbf{R} \delta\mathbf{q} \end{aligned} \tag{3.12}$$

donde en la segunda linea hemos hecho uso de la Ecuación 3.9. Ignorando el término de segundo orden $(\delta\mathbf{q})^T \mathbf{R} \delta\mathbf{q}$ y reemplazando en la Ecuación 3.10 tenemos:

$$\begin{aligned} \psi(\mathbf{q} + \delta\mathbf{q}) &= (\mathbf{q} + \delta\mathbf{q})^T \mathbf{R} (\mathbf{q} + \delta\mathbf{q}) \\ &= \psi(\mathbf{q}) + 2(\delta\mathbf{q})^T \mathbf{R} \mathbf{q} \end{aligned} \tag{3.13}$$

Por lo tanto, el uso de la Ecuación 3.11 en 3.12 implica que:

$$(\delta\mathbf{q})^T \mathbf{R} \mathbf{q} = 0 \tag{3.14}$$

Se observa que cualquier perturbación $\delta\mathbf{q}$ de $\mathbf{q}$ no es admisible. Mas precisamente, estamos restringidos a utilizar solo aquellas perturbaciones para las cuales la norma euclídea del vector $\delta\mathbf{q} + \mathbf{q}$ siga siendo unitario, esto es:

$$||\mathbf{q} + \delta\mathbf{q}|| = 1$$

o de manera equivalente,

$$(\mathbf{q} + \delta\mathbf{q})^T (\mathbf{q} + \delta\mathbf{q}) = 1$$

Por lo cual, observando la Ecuación 3.5, requerimos que para un primer orden en $\delta\mathbf{q}$,

$$(\delta\mathbf{q})^T\mathbf{q} = 0 \tag{3.15}$$

Esto significa que las perturbaciones $\delta\mathbf{q}$ deben ser ortogonales a $\mathbf{q}$ y por esto solo un cambio de dirección en $\mathbf{q}$ está permitido.

Por convención, los elementos del vector unitario $\mathbf{q}$ son adimensionales en un sentido físico. Si, por lo tanto, vamos a combinar las Ecuaciones 3.14 y 3.15, debemos introducir un factor de escala λ en la última ecuación con las mismas dimensiones que las entradas en la matriz de correlación $\mathbf{R}$. Podemos escribir entonces:

$$(\delta\mathbf{q})^T\mathbf{R}\mathbf{q} - \lambda(\delta\mathbf{q})^T\mathbf{q} = 0$$

o de forma equivalente,

$$(\delta\mathbf{q})^T(\mathbf{R}\mathbf{q} - \lambda\mathbf{q}) = 0 \tag{3.16}$$

Para que ésta última condición se cumpla, es necesario y suficiente que:

$$\mathbf{R}\mathbf{q} = \lambda\mathbf{q} \tag{3.17}$$

Esta es la ecuación que gobierna el vector unitario $\mathbf{q}$ para lo cual el estimador de la varianza $\psi(\mathbf{q})$ tiene valores extremos.

La Ecuación 3.17 es reconocida en álgebra lineal como el *problema de autovalores*. Este sistema tiene solución no trivial ($\mathbf{q} \neq 0$) solo para valores especiales de λ llamados *autovalores* de la matriz de correlación $\mathbf{R}$. Los valores asociados de $\mathbf{q}$ son llamados *autovectores*. Una matriz de correlación está caracterizada por autovalores reales y no negativos. Los autovectores asociados son únicos asumiendo que los autovalores sean distintos. Sean los autovalores de la matriz de correlación $\mathbf{R}$ de $m \times m$ denotados por $\lambda_1, \lambda_2, \ldots, \lambda_m$, y sus autovectores asociados denotados por $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m$, respectivamente.

Podemos escribir entonces:

$$\mathbf{R}\mathbf{q}_j = \lambda_j\mathbf{q}_j, \qquad j = 1, 2, \ldots, m, \tag{3.18}$$

Ordenamos los autovalores de forma decreciente:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_j > \ldots > \lambda_m \tag{3.19}$$

entonces $\lambda_1 = \lambda_{max}$. Los autovectores asociados para construir la matriz de $m \times m$ son:

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_j, \ldots, \mathbf{q}_m] \tag{3.20}$$

Podemos combinar el conjunto de las m Ecuaciones representadas en 3.18 en una sola Ecuación:

$$\mathbf{RQ} = \mathbf{Q\Lambda} \tag{3.21}$$

donde $\mathbf{\Lambda}$ es una matriz diagonal definida por los autovalores de la matriz $\mathbf{R}$:

$$\mathbf{\Lambda} = diag[\lambda_1, \lambda_2, \ldots, \lambda_j, \ldots, \lambda_m] \tag{3.22}$$

La matriz $\mathbf{Q}$ es una matriz unitaria ortogonal en el sentido que los vectores columna satisfacen la condición de ortonormalidad:

$$\mathbf{q}_i{}^T \mathbf{q}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{3.23}$$

La ecuación anterior requiere autovalores distintos. De forma análoga, podemos escribir:

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$$

donde se deduce que la inversa de la matriz $\mathbf{Q}$ es igual a su transpuesta:

$$\mathbf{Q}^T = \mathbf{Q}^{-1} \tag{3.24}$$

Esto indica que podemos reescribir la Ecuación 3.21 es una forma conocida como *transformación de similaridad ortogonal:*

$$\mathbf{Q}^T\mathbf{RQ} = \mathbf{\Lambda} \tag{3.25}$$

o en su forma expandida:

$$\mathbf{q}_i{}^T \mathbf{R}\mathbf{q}_k = \begin{cases} \lambda_j, & k = j \\ 0, & k \neq j \end{cases} \tag{3.26}$$

La transformación ortogonal de similaridad de la Ecuación 3.25 transforma la matriz de correlación $\mathbf{R}$ en una matriz diagonal de autovalores. $\mathbf{R}$ puede ser expresada en termino de sus autovalores y autovectores como:

$$\mathbf{R} = \sum_{i=1}^{m} \lambda_i \mathbf{q}_i \mathbf{q}_i^T \tag{3.27}$$

lo que es llamado *teorema espectral*. El producto externo $\mathbf{q}_i\mathbf{q}_i^T$ es de rango 1 para todo i.

Las Ecuaciones 3.25 y 3.27 son dos representaciones equivalentes de la *autodescomposición* de la matriz de correlación $\mathbf{R}$.

PCA y la autodescomposición de la matriz $\mathbf{R}$ son básicamente lo mismo pero observan el problema de diferente perspectiva. La equivalencia proviene de las Ecuaciones 3.10 y 3.27 donde se observa que el estimador de la varianza y los autovalores son en efecto iguales:

$$\psi(\mathbf{q}_j) = \lambda_j, \qquad j = 1, 2, \ldots, m \tag{3.28}$$

En resumen, hemos encontrado dos importantes hallazgos que refieren al conjunto de autovalores del análisis de componentes principales:

- Los autovectores de la matriz de correlación $\mathbf{R}$ pertenecientes al vector aleatorio $\mathcal{X}$ con media cero definen los vectores unitarios $\mathbf{q}_j$, representando las direcciones principales a lo largo de las cuales el estimador $\psi(\mathbf{q}_j)$ tiene sus valores extremos.
- Los autovalores asociados definen los valores extremos del estimador de la varianza $\psi(\mathbf{u}_j)$

3.2.2. Representación de los datos

Sea $\mathbf{x}$ un vector de datos que representa una realización del vector aleatorio $\mathcal{X}$.

Con m soluciones posibles para el vector unitario $\mathbf{q}$, encontramos que existen m proyecciones posibles del vector de datos $\mathbf{x}$ para ser consideradas. Específicamente de la Ecuación 3.4 notamos que:

$$a_j = \mathbf{q}_j^T \mathbf{x} = \mathbf{x}^T \mathbf{q}_j, \qquad j = 1, 2, \ldots, m \tag{3.29}$$

donde los a_j son las proyecciones de $\mathbf{x}$ en las direcciones principales representadas por los vectores unitarios $\mathbf{u}_j$. Los a_j son las llamadas *componentes principales* y tienen las mismas dimensiones físicas que el vector de datos $\mathbf{x}$. La fórmula en la Ecuación 3.29 puede ser vista como una de *análisis*.

Para reconstruir exactamente el vector de datos originales $\mathbf{x}$ a partir de las proyecciones a_j, procedemos de la siguiente manera. En primer lugar combinamos el conjunto de proyecciones $\{a_j | j = 1, 2, \ldots, m\}$ en un simple vector, esto es:

$$\begin{aligned} \mathbf{a} &= [a_1, a_2, \ldots, a_m]^T \\ &= [\mathbf{x}^T \mathbf{q}_1, \mathbf{x}^T \mathbf{q}_2, \ldots, \mathbf{x}^T \mathbf{q}_m]^T \\ &= \mathbf{Q}^T \mathbf{x} \end{aligned} \tag{3.30}$$

Luego, premultiplicamos ambos lados de la ecuación anterior por la matriz $\mathbf{Q}$, y luego utilizamos la relación de la Ecuación 3.24. En consecuencia, el vector de datos original $\mathbf{x}$

puede ser reconstruido de la siguiente manera:

$$\begin{aligned} \mathbf{x} &= \mathbf{Q}\mathbf{a} \\ &= \sum_{j=1}^{m} a_j \mathbf{q}_j \end{aligned} \tag{3.31}$$

la cual puede verse como la fórmula de *síntesis*. En este sentido, los vectores unitarios $\mathbf{q}_j$ representan una base en el espacio de los datos. En efecto, la Ecuación 3.31 no es más que una transformación de coordenadas, según la cual un punto $\mathbf{x}$ en el espacio de los datos es transformado a su correspondiente punto $\mathbf{a}$ en el espacio de características.

3.2.3. Reducción de dimensionalidad

Desde la perspectiva del reconocimiento de patrones, el valor práctico de PCA es que provee una efectiva técnica para reducir dimensionalidad. Particularmente, podemos reducir el numero de características necesarias para lograr una efectiva representación de los datos descartando aquellas combinaciones lineales de la Ecuación 3.31 que tienen pequeña varianza y retener solo aquellos términos que tienen varianza de mayor magnitud. Sean $\lambda_1, \lambda_2, \ldots, \lambda_l$ los l autovalores de mayor magnitud de la matriz de correlación $\mathbf{R}$. Podemos aproximar el vector de datos $\mathbf{x}$ *truncando* la expansión de la Ecuación 3.31 luego de los l términos:

$$\begin{aligned} \hat{\mathbf{x}} &= \sum_{j=1}^{l} a_j \mathbf{q}_j \\ &= [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_l] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_l \end{bmatrix}, \qquad l \leqslant m \end{aligned} \tag{3.32}$$

Dado el vector de datos original $\mathbf{x}$, podemos usar la Ecuación 3.29 para calcular el conjunto de componentes principales de la ecuación anterior de la siguiente manera:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \\ \mathbf{q}_l^T \end{bmatrix} \mathbf{x}, \qquad l \leqslant m \tag{3.33}$$

La proyección lineal de la Ecuación 3.33 de $\mathbb{R}^m$ a $\mathbb{R}^l$ representa el *codificador*. De manera análoga, la proyección lineal de la Ecuación 3.32 de $\mathbb{R}^l$ a $\mathbb{R}^m$ representa el *decodificador* para la aproximación de reconstrucción del vector original $\mathbf{x}$. Se observa que los autovalores de mayor magnitud $\lambda_1, \lambda_2, \ldots, \lambda_l$ no influyen en los cálculos de las Ecuaciones.

3.32 y 3.33; estos simplemente determinan el número de componentes principales usados para codificar y decodificar, respectivamente.

El *vector de error de aproximación* $\mathbf{e}$ es igual a la diferencia entre el vector original $\mathbf{x}$ y el vector aproximado $\hat{\mathbf{x}}$, esto es:

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}} \tag{3.34}$$

Substituyendo las Ecuaciones 3.31 y 3.32 en 3.34 tenemos:

$$\mathbf{e} = \sum_{j=l+1}^{m} \mathbf{a}_i \mathbf{q}_j \tag{3.35}$$

El vector de error $\mathbf{e}$ es *ortogonal* al vector de aproximación $\hat{\mathbf{x}}$. En otras palabras, el producto interno de $\hat{\mathbf{x}}$ y $\mathbf{e}$ es cero.

La varianza total de m componentes del vector $\mathbf{x}$ es, por la Ecuación 3.10 y la primer linea de 3.26:

$$\sum_{j=1}^{m} \sigma_j^2 = \sum_{j=1}^{m} \lambda_j \tag{3.36}$$

donde σ_j^2 es la varianza de la j-ésima componente principal a_j. La varianza total de los l elementos del vector aproximante $\hat{\mathbf{x}}$ es:

$$\sum_{j=1}^{l} \sigma_j^2 = \sum_{j=1}^{l} \lambda_j \tag{3.37}$$

La varianza total de los $(l-m)$ elementos en el vector de error de aproximación $\mathbf{x} - \hat{\mathbf{x}}$ es entonces:

$$\sum_{j=l+1}^{m} \sigma_j^2 = \sum_{j=l+1}^{m} \lambda_j \tag{3.38}$$

Los autovalores $\lambda_{l+1}, \ldots, \lambda_m$ son los $(m-l)$ mas pequeños de la matriz de correlación $\mathbf{R}$. Estos corresponden a los términos descartados en la expansión de la Ecuación 3.32 utilizada para construir el vector $\hat{\mathbf{x}}$. Cuanto más cercanos a cero estén todos estos autovalores, más efectiva será la reducción de dimensionalidad preservando el contenido de información del vector de entrada original. Por lo tanto, para reducir dimensionalidad de ciertos datos de entrada, *se calculan los autovalores y autovectores de la matriz de correlación del vector de entrada, y luego se proyectan los datos de entrada de forma ortogonal en el subespacio generado por los autovectores asociados a los autovalores dominantes.*

3.3. Redes Neuronales

Es posible entender el cerebro humano como un sistema de procesamiento de información muy complejo, altamente no lineal y con un elevado nivel de paralelismo. El mismo tiene la capacidad de organizar sus componentes estructurales, llamadas *neuronas*, de manera de realizar cálculos de distintos tipos (reconocimiento de patrones, percepción, control motor) muchas veces más rápido que cualquier computadora existente en la actualidad. Es en este sentido que surge el interés en desarrollar redes neuronales artificiales que simulen de mejor manera el procesamiento del cerebro humano.

Haykin [47] define a una red neuronal como:

> Una red neuronal es un procesador masivamente paralelo hecho de unidades de procesamiento simples que tiene una propensión natural para almacenar conocimiento proveniente de la experiencia y hacerlo disponible para su uso. Se asemeja al cerebro en dos aspectos:
>
> 1. El conocimiento es adquirido por la red de su ambiente mediante un proceso de aprendizaje.
> 2. Las fortalezas de las conexiones interneuronales, concidas como pesos sinápticos, son usadas para almacenar el conocimiento adquirido.

El procedimiento usado para realizar el proceso de aprendizaje se llama *algoritmo de aprendizaje*, cuya función es modificar los pesos sinápticos de manera de alcanzar un objetivo de diseño deseado. El algoritmo más utilizado es el *algoritmo de retropropagación*.

3.3.1. Perceptrón simple

Una *neurona* es una unidad de procesamiento simple que es fundamental para la operación de una red neuronal. El diagrama en bloque de la Figura 3.2 muestra el *modelo* de una neurona o *perceptrón simple* que es la base para el diseño de redes neuronales artificiales. Aquí se identifican tres elementos básicos del modelo neuronal:

1. Un conjunto de sinapsis o *enlaces de conexión* caracterizados por sus *pesos*. Específicamente, una señal x_j en la entrada sináptica j conectada a la neurona k es multiplicada por el peso sináptico w_{kj}. A diferencia de la sinapsis cerebral, un peso sináptico artificial puede estar en un rango que incluye valores negativos.

2. Una *sumatoria* que suma las señales de entrada, pesadas por sus correspondientes pesos sinápticos. Esta operación constituye una *combinación lineal*.

3. Una *función de activación* $\varphi(\cdot)$ que limita la amplitud de la salida de la neurona. Típicamente, el rango de amplitud de salida de una neurona pertenece a un intervalo cerrado $[0, 1]$ o $[-1, 1]$.

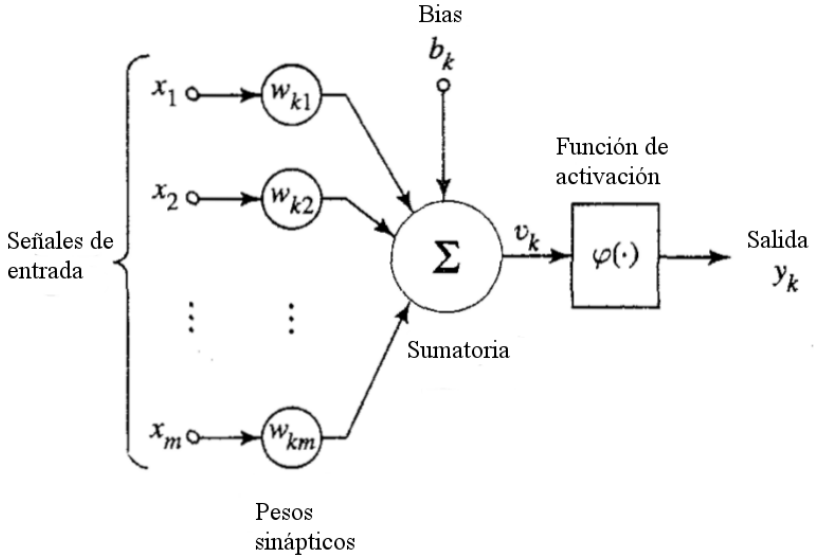

Figura 3.2: Modelo no lineal de una neurona [47].

El modelo neuronal que se muestra en la Figura 3.2 incluye un *bias* b_k, que tiene el efecto de incrementar o decrementar la salida de la neurona, dependiendo de si es positivo o negativo.

Matemáticamente, se puede describir un perceptrón simple k con el siguiente par de ecuaciones:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{3.39}$$

y

$$y_k = \varphi(u_k + b_k) \tag{3.40}$$

donde $x_1, x_2, \ldots, x_m$ son las señales de entrada; $w_{k1}, w_{k2}, \ldots, w_{km}$ son los pesos sinápticos de la neurona k; u_k es la salida de la combinación lineal de las entradas; b_k es el bias; $\varphi(\cdot)$ la función de activación e y_k la señal de salida de la neurona. El uso del bias tiene el efecto de aplicar una *transformación afín* a la salida u_k, como se ve en:

$$v_k = u_k + b_k \tag{3.41}$$

Dependiendo de un bias positivo o negativo, la relación entre el *potencial de activación* v_k de la neurona k y la salida de la combinación lineal u_k varía como se muestra en la Figura 3.3, para el caso de $m = 2$. Debido a esta transformación, la recta v_k frente a u_k ya no pasa por el origen. Se pueden combinar las Ecuaciones 3.39 a 3.41 para incluir al bias en la sumatoria, de la siguiente manera:

$$v_k = \sum_{j=0}^{m} w_{kj} x_j \tag{3.42}$$

y

$$y_k = \varphi(v_k) \tag{3.43}$$

con

$$x_0 = \pm 1 \tag{3.44}$$

y

$$w_{k0} = b_k \tag{3.45}$$

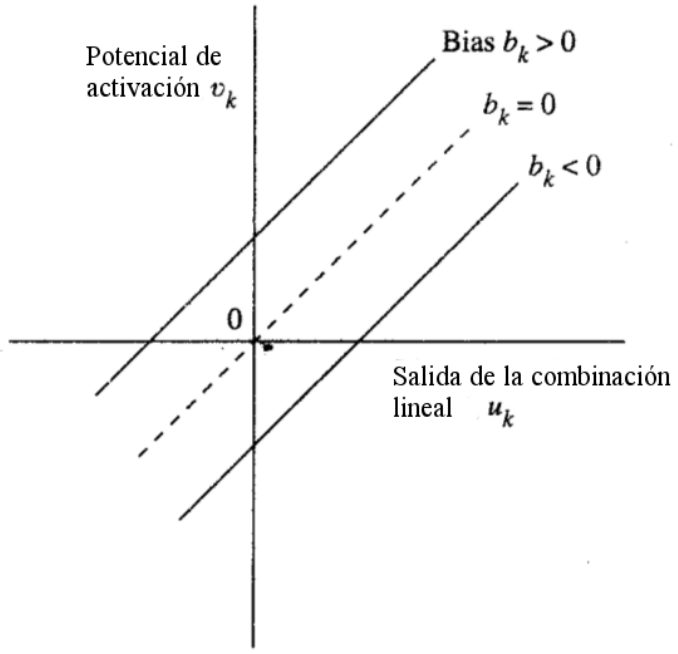

Figura 3.3: Transformación afín producida por la presencia de un *bias* [47].

Tipos de funciones de activación

La función de activación $\varphi(\cdot)$ define la salida de la neurona en función del potencial de activación v. Hay tres tipos básicos de funciones de activación [47]:

1. *Función escalón.* Con esta función, la salida de la neurona es 0 si v es negativo, o 1 si v es positivo. Esto se llama un comportamiento *todo o nada.*

$$\varphi(v) = \begin{cases} 1 & \text{si } v \geqslant 0 \\ 0 & \text{si } v < 0 \end{cases} \tag{3.46}$$

2. *Función lineal a trozos.* Esta función, definida por

$$\varphi(v) = \begin{cases} 1 & \text{si } v \geqslant +\frac{1}{2} \\ v & \text{si } +\frac{1}{2} > v > -\frac{1}{2} \\ 0 & \text{si } v \leqslant -\frac{1}{2} \end{cases} \tag{3.47}$$

es una aproximación a un comportamiento no lineal.

3. *Función sigmoidea.* Típicamente es la función de activación más utilizada. Se define como una función estrictamente creciente que tiene comportamiento balanceado entre lineal y no lineal. Un ejemplo de función sigmoidea es la *función logística*, definida por:

$$\varphi(v) = \frac{1}{1 + e^{-av}} \tag{3.48}$$

donde a es el parámetro que regula la *pendiente*. Variando este parámetro se obtienen distintas funciones sigomoideas. En el límite, cuando la pendiente tiende a infinito, la función sigmoidea se convierte en una función escalón. Esta función es diferenciable, mientras que la función escalón no lo es, y ésta es una característica imprescindible para el algoritmo de retropropagación, como explicaremos más adelante.

En la Figura 3.4 se pueden observar los tres tipos de funciones de activación.

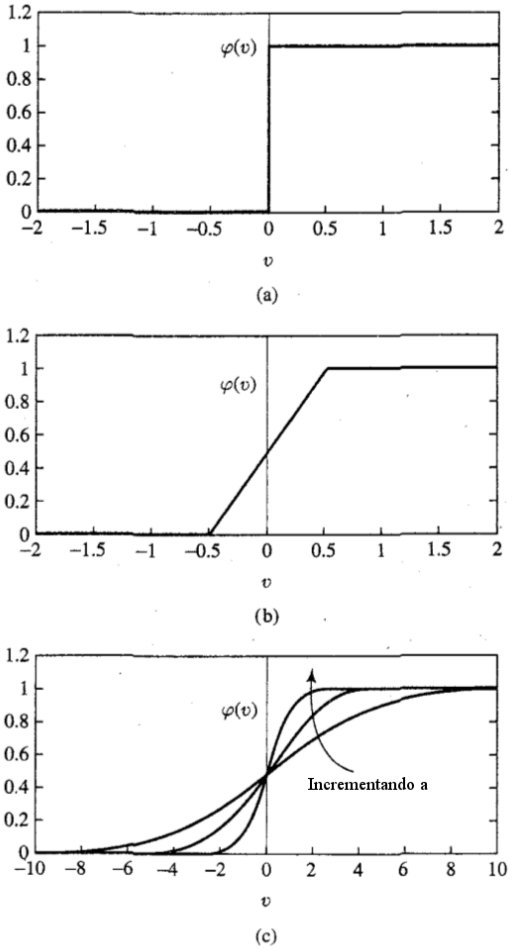

Figura 3.4: (a)Función escalón. (b) Función lineal a trozos. (c) Función sigmoidea variando parámetro a [47].

3.3.2. Perceptrón multicapa

Existen muchos tipos de arquitecturas neuronales, pero aquí nos concentraremos en la que es conocida como Perceptrón Multicapa, del inglés *Multilayer Perceptron* (MLP).

En este tipo de arquitectura, las neuronas están dispuestas en forma de capas. Un perceptrón multicapa se caracteriza por tener al menos una capa *oculta*, cuyas unidades de procesamiento son llamadas *neuronas ocultas*. La función de las capas ocultas es intervenir entre la *capa de entrada* que adquiere la información del exterior, y la *capa de salida*, que da la respuesta de la red neuronal. Esta dimensión extra da la habilidad a la red de obtener estadísticas de alto orden a partir de las entradas [47].

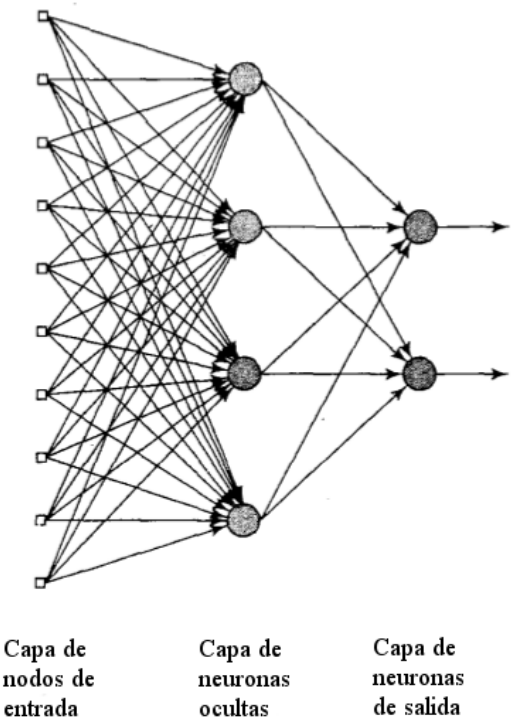

Figura 3.5: Perceptrón multicapa totalmente conectado de alimentación hacia adelante [47].

El perceptrón multicapa de la Figura 3.5 se denomina una red 10+4+2 debido a que tiene 10 neuronas en la capa de entrada, 4 neuronas en la capa oculta y 2 neuronas en la capa de salida. Generalizando, una red de n_i neuronas en la capa de entrada, n_{h1} neuronas en la primer capa oculta, n_{h2} neuronas en la segunda capa oculta y n_o neuronas en la capa de salida, se denomina una red $n_i + n_{h1} + n_{h2} + n_o$. Una red como la de la Figura 3.5 se dice que está *totalmente conectada*, en el sentido que cada nodo en cada capa de la red está conectado con todos los nodos de la capa siguiente. Si esto no fuera así, la red se llamaría *parcialmente conectada*.

3.3.3. Proceso de aprendizaje

La característica principal de una red neuronal es su capacidad de *aprender* de su entorno y de *mejorar* su rendimiento a través del aprendizaje. Una red neuronal aprende sobre su entorno a través un proceso interactivo de ajustes aplicados a sus pesos sinápticos y niveles de bias. Idealmente, la red conoce más sobre su ambiente luego de cada iteración del proceso de aprendizaje.

Haykin define aprendizaje como un proceso mediante el cual los parámetros libres de la red neuronal son adaptados a través de un proceso de estimulación por el ambiente en el que la red está embebida. El tipo de aprendizaje está dado por la manera en que se da este cambio de los parámetros [47].

Según esta definición, hay 3 pasos principales en el proceso de aprendizaje:

1. La red neuronal es *estimulada* por el ambiente.
2. La red neuronal *sufre cambios* en sus parámetros libres como resultado de esta estimulación.
3. La red neuronal responde al ambiente *de una nueva forma* debido a estos cambios que sufrió en su estructura interna.

Un conjunto de reglas bien definidas para la solución del problema de aprendizaje es llamada un *algoritmo de aprendizaje*. Hay una variedad de algoritmos de aprendizaje, que difieren principalmente en la forma en la que se ajustan los pesos sinápticos. Otro factor a tener en cuenta es la forma en la que la red neuronal se relaciona con el ambiente. En la sección anterior describimos brevemente un perceptrón multicapa, que tiene una forma particular de relacionarse con el ambiente, pero éste no es el único tipo de red neuronal.

Antes de repasar el algoritmo de aprendizaje más utilizado para redes neuronales del tipo perceptrón multicapa, es conveniente mencionar que existen 5 reglas básicas de aprendizaje: aprendizaje de *corrección de error*, aprendizaje *basado en memoria*, aprendizaje *Hebbiano*, aprendizaje *competitivo* y aprendizaje de *Boltzmann*. El aprendizaje de corrección de error ajusta los pesos de manera de minimizar una función de error entre la salida de la red neuronal y una salida esperada. El resto de las reglas describen otros tipos de aprendizaje que no son utilizados en este trabajo.

Por otro lado, existen dos principales paradigmas de aprendizaje: aprendizaje *supervisado* y aprendizaje *no supervisado*. El primero de los mismos se refiere a que el algoritmo de aprendizaje tiene en cuenta información del ambiente, dada por un conjunto de ejemplos del tipo *entrada-salida*. En el segundo paradigma, no se cuenta con información del ambiente *a priori*, y la red debe aprender *sin supervisión*.

El algoritmo de retropropagación es un algoritmo de corrección de error, que cae dentro del paradigma de aprendizaje supervisado.

3.3.4. Algoritmo de retropropagación

Como dijimos anteriormente, una red neuronal multicapa de propagación hacia adelante consiste en un conjunto de unidades sensitivas que captan información del ambiente y constituyen la *capa de entrada*, una o más *capas ocultas* que realizan el procesamiento de esta información y una *capa de salida*. La señal de entrada se propaga en una dirección (hacia adelante) por la red capa a capa.

El algoritmo de *retropropagación del error* está basado en la regla de aprendizaje de corrección del error. Básicamente, el aprendizaje de retropropagación del error consiste en dos pasadas a través de las diferentes capas de la red: una hacia adelante y otra hacia atrás. En la *pasada hacia adelante*, un patrón de actividad (vector de entrada) es aplicado a las unidades sensoriales de la red y su efecto se propaga hacia adelante capa a capa. Finalmente, se producen un conjunto de salidas que constituyen la respuesta de la red a la excitación de entrada. Durante la pasada hacia adelante los pesos sinápticos de la red se mantienen *fijos*. Durante la *pasada hacia atrás*, en cambio, los pesos son *ajustados* de acuerdo con una regla de corrección de error. Específicamente, la respuesta real de la red es restada de una respuesta deseada (objetivo) para producir una *señal de error*. Esta señal de error se propaga hacia atrás en la dirección opuesta a la dirección de las conexiones sinápticas. De aquí el nombre del algoritmo de "propagación hacia atrás del error".

A continuación se hará un resumen del funcionamiento del algoritmo de retropropagación. Para detalles de la derivación de las ecuaciones se puede consultar la extensiva descripción en Haykin [47].

Existen dos modos principales de entrenamiento: *secuencial* y *por lotes*. En el primer modo, los pesos son actualizados para cada ejemplo de entrenamiento $\{x(n), d(n)\}$, donde $x(n)$ corresponde al patrón de actividad de entrada y $d(n)$ a la salida deseada (objetivo). En el segundo, los pesos se actualizan luego de la presentación de todos los ejemplos de entrenamiento, lo que se denomina una *época* de entrenamiento.

Para el primer modo de entrenamiento, el algoritmo itera a través del ejemplo de entrenamiento $\{x(n), d(n)\}_{n=1}^{N}$ como sigue:

1. *Inicialización*. Asumiendo que no hay información previa disponible, se eligen los pesos sinápticos y los umbrales aleatoriamente de una distribución uniforme con media cero y cuya varianza se elige para que la desviación estándar de los potenciales de activación de las neuronas caigan en la transición entre la parte lineal y saturada de la función de activación sigmoidea. Dependiendo de la función de activación utilizada, generalmente los pesos se eligen para que sean aleatorios en el rango $[0, 1]$ o en el rango $[-1, 1]$.

2. *Presentación de los ejemplos de entrenamiento*. Se le presenta a la red una época de los ejemplos de entrenamiento. Para cada ejemplo en el conjunto se ejecutan los pasos de propagación hacia adelante y hacia atrás descriptos en los pasos 3 y 4, respectivamente.

3. *Propagación hacia adelante.* Denotamos un ejemplo de entrenamiento en la época como $(\mathbf{x}(n), \mathbf{d}(n))$. El vector de entrada $\mathbf{x}(n)$ se aplica a la capa de entrada de la red y la salida deseada $\mathbf{d}(n)$ se le presenta a la capa de salida. Se computan los potenciales de activación de la red procediendo hacia adelante, capa a capa. El potencial de activación $v_j^{(l)}(n)$ para una neurona j en una capa l es

$$v_j^{(l)}(n) = \sum_{i=0}^{m_0} w_{ji}^{(l)}(n) y_i^{(l-1)}(n) \tag{3.49}$$

donde $y_i^{(l-1)}(n)$ es la señal de salida de la neurona i en la capa previa $l-1$ en la iteración n y $w_{ji}^{(l)}(n)$ es el peso sináptico de la neurona j en la capa l que es alimentada por la neurona i en la capa $l-1$. Para $i=0$, tenemos $y_0^{(l-1)}(n) = +1$ y $w_{j0}^{(l)} = b_j^{(l)}(n)$ es el bias aplicado a la neurona j en la capa l. Asumiendo el uso de función de activación sigmoidea, la señal de salida de la neurona j en la capa l es

$$y_j^{(l)} = \varphi_j(v_j(n)) \tag{3.50}$$

Si la neurona j está en la primer capa oculta ($l=1$), decimos que

$$y_j^{(0)}(n) = x_j(n) \tag{3.51}$$

donde $x_j(n)$ es el j-ésimo del vector de entrada $\mathbf{x}(n)$. Si la neurona j está en la capa de salida ($l=L$, donde L es la *profundidad* de la red), decimos que

$$y_j(L) = o_j(n) \tag{3.52}$$

Se computa la señal de error

$$e_j(n) = d_j(n) - o_j(n) \tag{3.53}$$

donde $d_j(n)$ es el j-ésimo elemento del vector de salida deseada $\mathbf{d}(n)$.

4. *Propagación hacia atrás.* Se computan los gradientes locales (δs) de la red, definidos como

$$\delta_j^{(l)}(n) = \begin{cases} e_j^{(L)}(n)\varphi_j'(v_j^{(L)}(n)) & \text{para neurona } j \text{ en capa de salida } L \\ \varphi_j'(v_j^{(L)}(n)) \sum_k \delta_k^{(l+1)}(n) w_{kj}^{(l+1)}(n) & \text{para neurona } j \text{ en capa oculta } l \end{cases}$$

donde $\varphi_j'(\cdot)$ denota diferenciación respecto del argumento. Se ajustan los pesos sinápticos de la red de acuerdo con la regla delta generalizada:

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \eta[w_{ji}^{(l)}(n-1)] + \mu\delta_j^{(l)}(n) y_j^{(l-1)}(n) \tag{3.54}$$

donde μ es el parámetro de tasa de aprendizaje y η es el término de momento.

El algoritmo de retropropagación provee una "aproximación" a la trayectoria en el espacio de los pesos calculada por el método de gradiente descendente. Mientras más

pequeña la tasa de aprendizaje más pequeños serán los cambios a los pesos sinápticos de la red de una iteración a la siguiente y la trayectoria en el espacio de los pesos será más suave. Esta mejora, sin embargo, viene al costo de un aprendizaje de la red más lento. Si, por el contrario, la tasa de aprendizaje es demasiado grande con el objetivo de acelerar el aprendizaje de la red, los cambios en los pesos sinápticos pueden volverse inestables (por ejemplo, oscilatorios).

El término de momento tiene un *efecto acelerador* en las direcciones de la superficie de error que descienden establemente. En zonas de la superficie en las que hay oscilación en signo, el término de momento tiene un *efecto estabilizador*. El ajuste de pesos $\Delta w_{ji}(n)$ representa la suma de una serie de tiempo pesada exponencialmente. Para que la serie de tiempo sea convergente, el término de momento debe estar restringido al rango $0 \leqslant |\eta| < 1$. Cuando η es cero, el algoritmo de retropropagación opera sin término de momento.

5. *Iteración.* Se iteran la propagación hacia adelante y hacia atrás de los pasos 3 y 4 mediante la presentación de nuevas épocas de entrenamiento a la red hasta que se alcanza algún criterio de detención del algoritmo.

En general, no puede demostrarse la convergencia del algoritmo de retropropagación, y no hay un único criterio para detener su operación. Por lo tanto se utilizan criterios prácticos para la detención del ajuste de los pesos. Para formular estos criterios se piensa en función de las propiedades únicas de un *mínimo global* o *local* de la superficie de error.

Llamamos $\mathbf{w}^*$ al vector de pesos que denota un mínimo en la superficie de error, sea local o global. Una condición necesaria para que $\mathbf{w}^*$ sea un mínimo es que el vector gradiente $\mathbf{g}(\mathbf{w})$ (la derivada parcial de primer orden) de la superficie de error con respecto al vector de pesos $\mathbf{w}$ sea cero en $\mathbf{w} = \mathbf{w}^*$. De esta manera, se puede formular un criterio de detención del algoritmo que tenga en cuenta que la *norma euclídea del vector gradiente llegue a un umbral lo suficientemente pequeño*. Sin embargo, este criterio de convergencia genera tiempos de aprendizaje muy largos y tiene como desventaja la necesidad de calcular el gradiente $\mathbf{g}(\mathbf{w})$.

Otra propiedad de un mínimo que se puede utilizar es que la función de costo o medida del error $\xi_{av}(\mathbf{w})$ es estacionaria en el punto $\mathbf{w} = \mathbf{w}^*$. De esta manera, el algoritmo puede detener su operación cuando *la tasa de cambio absoluta del promedio del error cuadrático medio por época es suficientemente pequeña*. Generalmente se considera que esta tasa de cambio es lo suficientemente pequeña si cae en el rango de 0,1 a 1 % por época. Este método tiene la desventaja de que puede resultar en cortes prematuros del proceso de aprendizaje.

Un criterio más adecuado utilizado en conjunto con la técnica de validación cruzada es probar la capacidad de generalización de la red, y detener el entrenamiento cuando se alcanza el pico de generalización.

Capítulo 4

Método Propuesto

El método propuesto para el reconocimiento automático de emociones en contenido multimedia consta de dos etapas: la etapa de extracción de características y la etapa de clasificación de estas características en clases. Estas clases son una de las seis emociones básicas: ira, asco, miedo, alegría, tristeza y sorpresa.

En la Figura 4.1 se ve un diagrama simplificado del método propuesto.

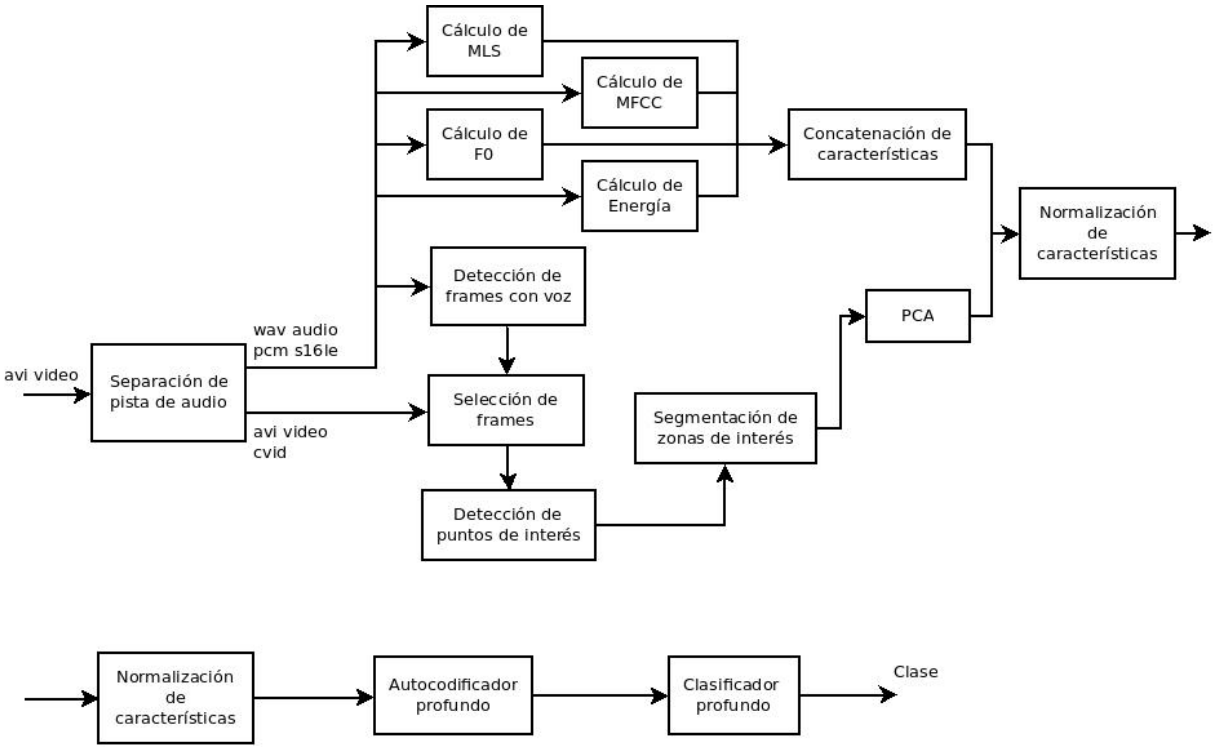

Figura 4.1: Diagrama descriptivo del método propuesto.

La primera etapa consiste en la separación de la pista audio de los archivos de video. La extracción de características de audio se basa en el trabajo de Albornoz et al. [4]. Las características de audio que se obtienen son MFCCs junto con sus derivadas primera

y segunda, coeficientes Media del Espectro Logarítmico, del inglés *Mean Log-Spectrum* (MLS), media y desvío de la energía de corta duración y media y desvío la frecuencia fundamental (F_0).

La extracción de características de video se realiza aplicando PCA sobre zonas de interés. Estas zonas de interés son la zona de la boca y de los ojos. Se seleccionaron estas zonas debido al alto nivel de información emocional que contienen [11].

Luego de la etapa de extracción de características se cuenta con un vector de características por cada video. La clasificación se realiza mediante autocodificadores profundos apilados. Se construyó una red neuronal profunda de 2 capas ocultas y 6 neuronas en la capa de salida.

Se optimizó la arquitectura utilizando la técnica de preentrenamiento por capas, antes de realizar retropropagación sobre la red entera [49]. Se optimizaron los siguientes parámetros: cantidad de neuronas en cada capa oculta, función de activación y parámetros de entrenamiento.

En la Sección 4.1 se detalla el método propuesto para la extracción de características, tanto de audio como de video.

En la Sección 4.2 se detalla el método propuesto para la clasificación. Primero se hace una introducción a la técnica de reducción de dimensionalidad de autocodificadores profundos, y luego se detalla el método adoptado.

4.1. Extracción de Características

La etapa de extracción de características presenta dos subetapas separadas: la extracción de características en audio y la extracción de características en video. Si bien el audio y el video se procesan de manera separada, en el procesamiento del video se aprovecha el hecho de contar con información de audio para detectar los *frames* donde hay mayor probabilidad de presencia de voz.

4.1.1. Extracción de características en audio

Para la extracción de características de audio se obtienen 3 tipos de características acústicas: características prosódicas, como son la energía y la frecuencia fundamental, características espectrales, como los MLS y características cepstrales como los MFCC. Esta elección de características está inspirada en el trabajo de Albornoz et al. [4].

En la Figura 4.2, se puede apreciar la variación del *pitch* estimado a lo largo de una frase de un ejemplo de la base de datos RML Emotion Database[1]. Además, se puede observar en la parte superior de la imagen la probabilidad que haya voz en la señal.

[1] `http://www.rml.ryerson.ca/rml-emotion-database.html`

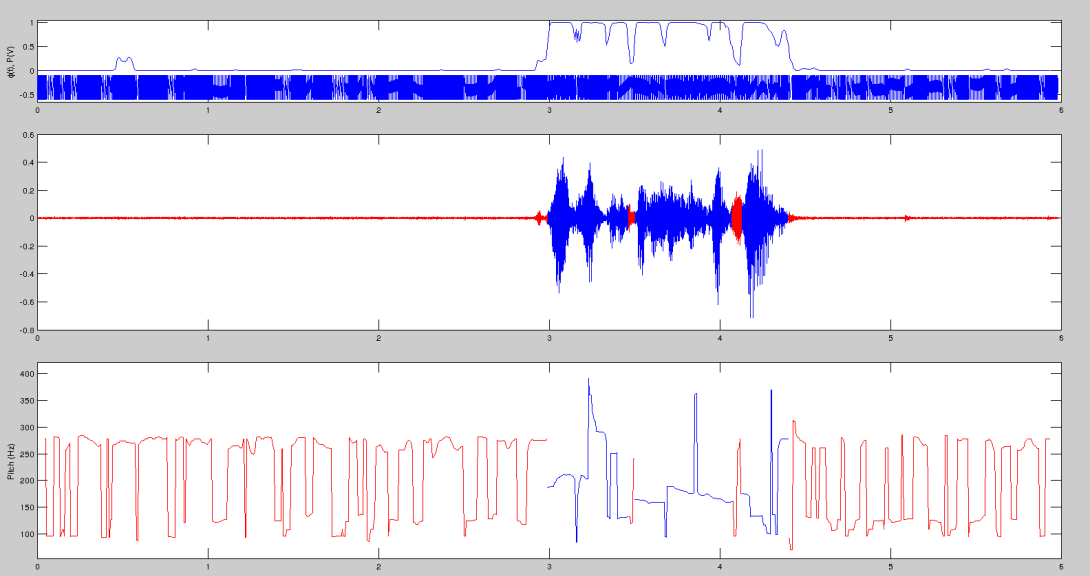

Figura 4.2: Probabilidad de que haya voz en la señal (arriba), forma de onda de la señal (medio) y estimación del pitch (abajo).

En este trabajo se calculan la media y el desvío estándar del *pitch* o frecuencia fundamental (F_0). La frecuencia fundamental es la frecuencia más baja del espectro de frecuencias de una onda periódica. Como la voz es una onda cuasi-periódica en los sonidos sonoros (con componente glótica) se puede estimar el *pitch*, que es una descripción subjetiva psico-acústica de la onda sonora en la que se ubica a la onda en una posición relativa en una escala relacionada a las frecuencias. Para realizar esto se utiliza el método PEFAC [45], que es un método de estimación del *pitch* robusto a altos niveles de ruido. Este algoritmo estima la frecuencia fundamental de cada *frame* mediante la convolución de su espectro de potencia en el dominio logarítmico con un filtro que suma las energías de las armónicas.

Para una fuente periódica con *pitch* f_0 y asumiendo ruido estacionario, la densidad espectral de potencia en el dominio logarítmico de las frecuencias está dada por:

$$Y(q) = \sum_{k=1}^{K} b_k \delta(q - \log k - \log f_0) + N(q) \tag{4.1}$$

donde $q = \log f$, b_k representa la potencia de la k-ésima armónica, $N(q)$ la densidad espectral de potencia del ruido, δ la función Delta de Dirac, y K el número de armónicas. En el dominio logarítmico de las frecuencias, la separación de las armónicas no depende de f_0 y su energía puede ser sumada mediante la convolución de $Y(q)$ con un filtro específico, cuya respuesta al impulso es:

$$h(q) = \sum_{k=1}^{K} \delta(q - \log k) \tag{4.2}$$

La convolución $Y(q) * h(q)$ va a resultar en un pico en $q_0 = \log f_0$, junto con otros picos adicionales correspondientes a los múltiplos racionales de f_0. En principio, el *pitch* f_0

puede ser obtenido a través del mayor pico en la salida del filtro. El algoritmo PEFAC modifica el filtro ideal definido en la Ecuación 4.2 debido a que tanto la señal como el ruido son no estacionarios.

En la Figura 4.3 se puede apreciar el *pitch* estimado superpuesto al espectograma de la señal.

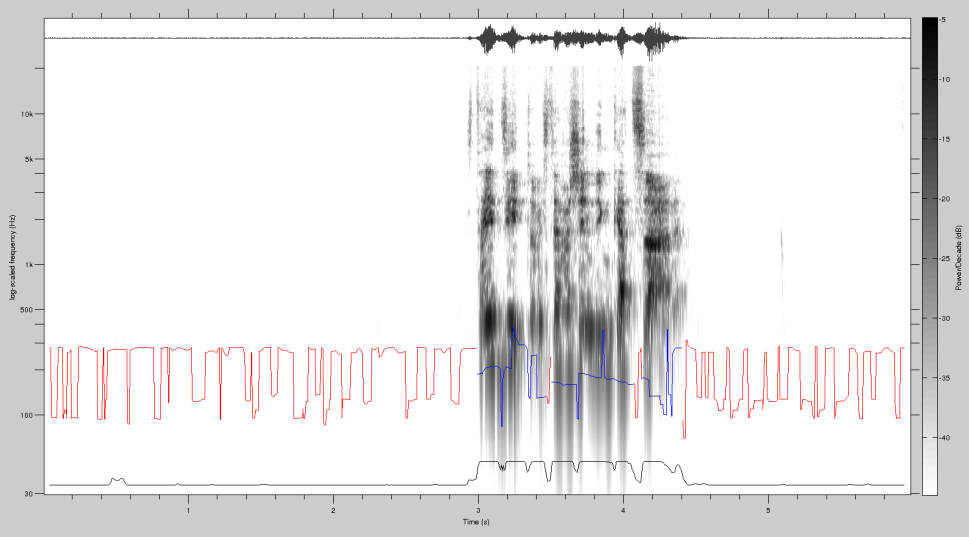

Figura 4.3: Espectograma con pitch superpuesto.

La energía de corta duración es un parámetro muy utilizado para el análisis de discurso hablado. Se define como:

$$E_n = \sum_{m=-\infty}^{m=+\infty} [x(m)w(n-m)]^2 \tag{4.3}$$

donde w es la ventana utilizada. Esta medida puede de alguna manera distinguir los segmentos de voz que presentan sonidos con componente glótica de los sonidos sordos, ya que los primeros presentan energía de corta duración de significativa mayor magnitud.

En la Figura 4.4 se puede apreciar la energía de corta duración para un ejemplo de la base de datos. Para mejor visibilidad se recortó la señal a la parte con voz.

La media del espectro logarítmico MLS se calcula para cada banda de frecuencias a lo largo del tiempo de la siguiente manera:

$$S(k) = \sum_{n=1}^{N} \log |(v(n,k))| \tag{4.4}$$

donde k es la banda frecuencial en la que se calcula, N es el número de *frames* en el ejemplo de video, y $v(n,k)$ es la Transformada Discreta de Fourier en el *frame* n. Se utilizaron 30 coeficientes MLS en el rango de frecuencias (0-1200Hz) dado que en [4] se verifica que estos son los que presentan mejores resultados de separacación de emociones.

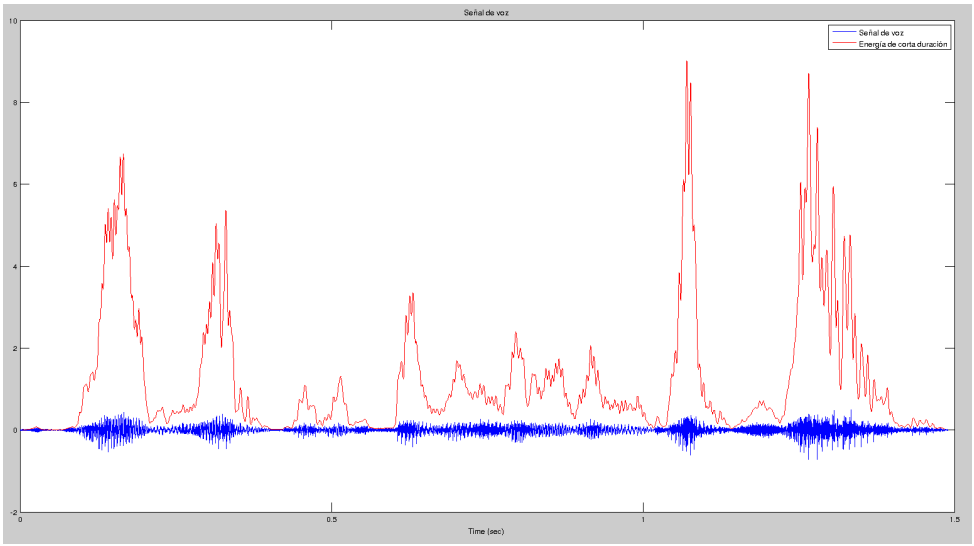

Figura 4.4: Señal de audio (azul) recortada al sector donde presenta voz, y energía de corta duración (rojo).

El vector de características de audio, que llamamos $\mathbf{fva}_{46}$, queda conformado entonces por:

- 12 medias de los MFCCs. Para el cálculo de estos se utilizan ventanas en el dominio del tiempo del tipo Hamming con un tamaño de 1024 *frames* de audio.
- 30 coeficientes MLS.
- La media y el desvío estándar de la energía de corta duración. Para realizar el cálculo de la energía de corta duración se utilizan ventanas de 10 ms. [25].
- La media y el desvío de la frecuencia fundamental (F_0). Para el cálculo se utilizan ventanas de 10 ms.

4.1.2. Extracción de características en video

Para la extracción de características en video se desarrollaron algoritmos para encontrar regiones de interés por *frame* de video. Estas regiones o zonas de interés corresponden a la zona de la boca y a la zona de los ojos del sujeto en el video. Se optó por seleccionar dichas regiones dado que son las que suponen mayor contenido emocional en las expresiones faciales [11]. Luego se aplica PCA para obtener un vector de 70 características por video.

A continuación se enumeran los pasos que se siguen para lograr la extracción de estas regiones en las imágenes de video:

1. Selección de frames en el video.

2. Detección de rostro en video.
3. Detección de *puntos de interés* o Puntos Faciales de Interés, del inglés *Facial Landmarks* (FL).
4. Segmentación de *zonas de interés* (ojos y boca).
5. Normalización de imágenes.
6. Creación de particiones para validación cruzada.
7. Aplicación de PCA.

El primer paso consiste en determinar qué *frames* de video se deben considerar para la extracción de características y la posterior clasificación. Para esto, se realizó una exhaustiva observación de cada uno de los ejemplos de la base de datos a utilizar y se constató que en cada uno de los videos los sujetos parten del reposo (ausencia de emoción), expresan una frase (con carga emocional) y vuelven al reposo. El objetivo entonces fue acotar el video al momento de la expresión emocional. Es por esto que se desarrolló un algoritmo que calcula en qué *frame* de video el sujeto comienza a hablar y en cuál termina de hablar, por lo que ese intervalo de fotogramas es el que se tiene en cuenta para la posterior clasificación.

Se desarrolló entonces una función que a partir de la pista de audio del video determina el instante de tiempo en que la persona comienza a experimentar una emoción y el instante final antes de volver al estado de reposo.

Se verificó que los sujetos de la base de datos pronuncian las frases con carga emocional durante un promedio de 47 *frames* de video. Se decidió usar esta cantidad de *frames* para todos los videos, de manera de simplificar la reducción de dimensionalidad. De esta manera, en los casos en los que la cantidad de *frames* con carga emocional era superior a 47 se descartaron los *frames* de los extremos. De la misma manera, y siguiendo el mismo razonamiento, en los casos en los que la emoción se mantiene por menos de 47 *frames*, se decidió repetir el *frame* central, que se considera es uno de los que más carga emocional presenta.

Una vez seleccionados los *frames* que presentan carga emocional, se procede a la segmentación de las zonas de interés. Para esto, primero es necesario localizar los puntos importantes en los rostros. Para realizar el paso 2 y 3 se utilizó la librería *flandmark* [99] la cual permite detectar rostros en imágenes y luego calcular los FL. Los FL son 8 puntos que permiten determinar con exactitud la localización del centro de la cara (ϵ_0) los ojos (ϵ_5 y ϵ_1 para ojo derecho y ϵ_2 y ϵ_6 para ojo izquierdo), la boca (ϵ_3 y ϵ_4) y la nariz (ϵ_7); como se ve en la Figura 4.5. Esta librería utiliza un clasificador basado en Modelo de Partes Deformables, del inglés *Deformable Part Models* (DPM) [39] y está documentado un alto nivel de detección de estos marcadores en el rostro [99]. En la Figura 4.6 se ve la localización de los 8 FL en un ejemplo de la base de datos.

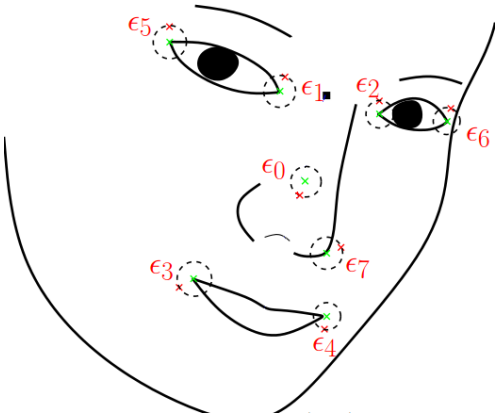

Figura 4.5: Muestra de los 8 FL. Imágen adaptada de [99].

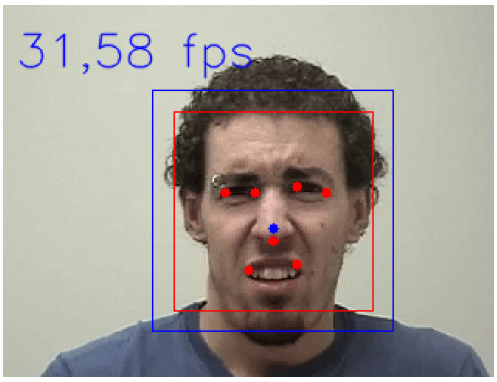

Figura 4.6: FL en un ejemplo de la base de datos.

Una vez determinados los puntos de interés, se procede a segmentar las zonas de interés. Se consideran ciertos márgenes de desvío para capturar no sólo información de la boca y los ojos, sino también información relevante con contenido emocional como lo son las cejas y pequeñas muecas alrededor de la boca [11]. En la Figura 4.7 se pueden observar las zonas de interés, en las que se aprecian las marcas de expresión características. Estas imágenes capturadas son guardadas en escala de grises con valores de [0,255].

El paso cinco se definió con el objetivo de simplificar la reducción de dimensionalidad. Cada región de interés (boca y ojos) queda fijada a un tamaño de 70 x 35 píxeles, lo que da un total de 4900 valores por frame de video.

El sexto paso consiste en generar las particiones de entrenamiento, validación y prueba para la validación cruzada, como se detallará en la Sección 4.2.3.

Figura 4.7: Muestra de las zonas de interés segmentadas.

El último paso consiste en aplicar PCA. Para cada una de las particiones, se generan las matrices de proyección con el conjunto de datos de entrenamiento, y luego se proyectan todos los vectores para obtener la representación reducida en el nuevo espacio. Es decir, se aplica PCA una vez por cada partición.

Para generar los nuevos espacios, se tuvo en cuenta retener al menos el 85 % de la varianza para cada una de las emociones, lo que significó fijar 70 componentes principales por espacio.

De esta manera, se genera el vector de características en video llamado $\mathbf{fvv}_{70}$ el cual posee cada una de las componentes principales antes mencionadas.

4.2. Clasificación

En esta sección se detallará el clasificador diseñado para lograr el reconocimiento de emociones en contenido multimodal. Este clasificador es un Perceptrón Multicapa, del inglés *Multilayer Perceptron* (MLP), pre-entrenado mediante la técnica de autocodificación profunda. La red neuronal se construye apilando autocodificadores, optimizando la arquitectura capa por capa, para obtener los mejores resultados de autocodificación y clasificación finales.

4.2.1. Autocodificadores profundos

Un autocodificador o autoencoder es una red neuronal con una arquitectura particular. El objetivo de los autocodificadores es aprender una representación comprimida de los datos con el fin de lograr reducir dimensionalidad en ellos.

En general, la estructura de un autoencoder es muy similar a la de un MLP; son redes no recurrentes y de propagación hacia adelante con una capa de entrada, una capa de salida, y una o mas capas ocultas que conectan estas primeras. La diferencia con un MLP es que la capa de salida tiene la misma cantidad de neuronas que la capa de entrada y en vez de entrenar la red con el objetivo de predecir una clase y a partir de un patrón x el autocodificador es entrenado para poder reconstruir el propio patrón de entrada x. Si se utiliza sólo una capa oculta y las neuronas tienen funciones de activación lineales, la solución óptima encontrada por el autocodificador está relacionada con PCA [12]. Sin embargo, si se utilizan funciones de activación sigmoideas y más cantidad de capas se puede aprovechar la no linealidad para encontrar una representación no lineal más eficiente [49].

Hinton [49] propuso utilizar una serie de autocodificadores apilados para realizar un pre-entrenamiento (capa a capa) de la red con el objetivo de llevar los pesos a valores cercanos de una buena solución y evitar así que la red se estanque en mínimos locales debidos a la naturaleza misma del algoritmo de retropropagación.

En el siguiente apartado se detalla como fue entrenado el autocodificador en este trabajo, y como éste se convierte en el clasificador profundo luego de haberle añadido la última capa.

4.2.2. Diseño de la Arquitectura

El clasificador es una red neuronal de (**i**) neuronas en la capa de entrada, dadas por la cantidad de características obtenidas en audio y en video; y 6 neuronas en la capa de salida (**o**), dadas por una de seis emociones presentes. Este clasificador, denominado *clasificador profundo*, es el resultado la concatenación del *autocodificador profundo* con la capa de salida, como mostraremos mas adelante.

La arquitectura de la red neuronal se elige optimizando los parámetros de cada capa y la cantidad de capas ocultas mediante la técnica de autocodificadores apilados [49] como mencionamos anteriormente. En este trabajo se utilizaron 2 capas ocultas ($\mathbf{h}_1$ y $\mathbf{h}_2$) por lo que los parámetros a optimizar fueron:

- cantidad de neuronas en cada capa (n_{h1}, n_{h2}),
- función de activación de cada capa y,
- parámetros de entrenamiento del algoritmo de retropropagación con término de momento.

Comenzaremos explicando entonces como se da el entrenamiento del autocodificador profundo paso a paso. Para el entrenamiento de cada red, se selecciona la de mejor rendimiento utilizando el criterio de corte por el Error Cuadrático Medio, del inglés *Mean Squared Error* (MSE) de la partición de validación como se detalla en la siguiente sección.

Etapa 1

El primer paso consiste en realizar el pre-entrenamiento de la primer capa; esto es, la inicialización de los pesos entre la capa $\mathbf{i}$ y la capa $\mathbf{h}_1$. Para lograr esto, se diseña el primer autocodificador como se observa en la Figura 4.8. Llamamos a este autocodificador $\mathbf{i} + \mathbf{h}_1 + \mathbf{i}$.

Etapa 2

Una vez finalizada la Etapa 1, se toma la mejor red resultante de esta etapa y se elimina la capa de salida resultando así el autocodificador $\mathbf{i} + \mathbf{h}_1$, como se muestra en la Figura 4.9. Luego se pasan cada una de las particiones de entrenamiento, validación y prueba por esta red para generar los patrones de entrenamiento de la etapa siguiente.

Etapa 3

Se construye el autocodificador $\mathbf{h_1} + \mathbf{h_2} + \mathbf{h_1}$ y se entrena con los patrones generados en el paso anterior. Este autocodificador es análogo al de la Etapa 1 y se puede observar su estructura en la Figura 4.10.

Etapa 4

Nuevamente se selecciona la mejor red del paso anterior en cada una de las particiones y se genera la red $\mathbf{h}_1 + \mathbf{h}_2$, como se muestra en la Figura 4.11.

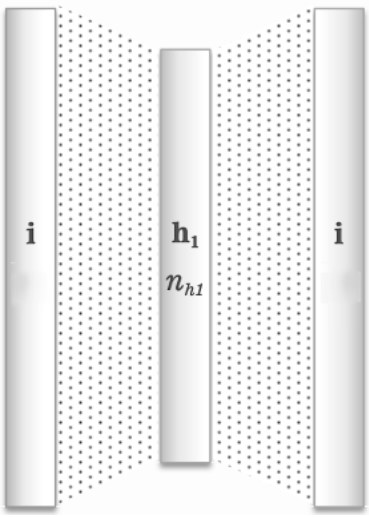

Figura 4.8: Autocodificador de la Etapa 1.

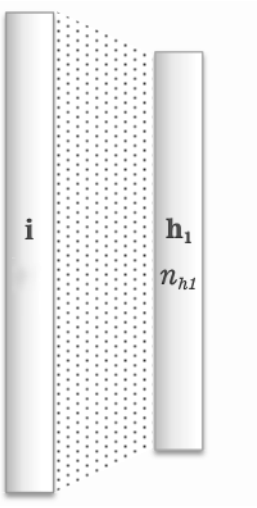

Figura 4.9: Red resultante de la Etapa 2.

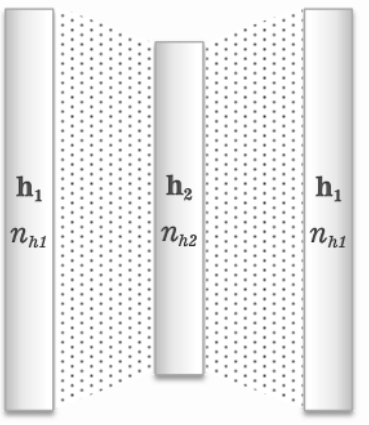

Figura 4.10: Autocodificador de la Etapa 3.

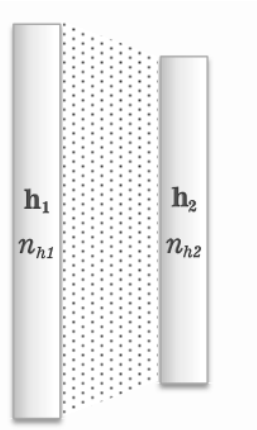

Figura 4.11: Red resultante de la Etapa 4.

Etapa 5

Esta etapa tiene como objetivo crear el autocodificador profundo. En primer lugar, se unen las redes de las Etapas 2 y 4. Es decir, se concatenan las redes $\mathbf{i} + \mathbf{h}_1$ y $\mathbf{h}_1 + \mathbf{h}_2$ para formar la red $\mathbf{i} + \mathbf{h}_1 + \mathbf{h}_2$; se utilizan conexiones 1 a 1 entre las capas $\mathbf{h}_1$ (pesos unitarios que no se modifican). Luego, se hace una inversión de esta red y se la denota como $\mathbf{h}'_2 + \mathbf{h}'_1 + \mathbf{i}'$ (se utiliza " $'$ " sólo con el fin de distinguir gráficamente las capas originales de las invertidas). Finalmente se unen estas dos partes para formar el autocodificador profundo $\mathbf{i} + \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}'_1 + \mathbf{i}'$.

En la Figura 4.12 se puede observar como queda definido el mismo. Nuevamente se pasan los patrones de entrenamiento y el autocodificador profundo se entrena de forma análoga a los dos anteriormente entrenados. Una vez completado el entrenamiento, se selecciona el que mejor rendimiento logró.

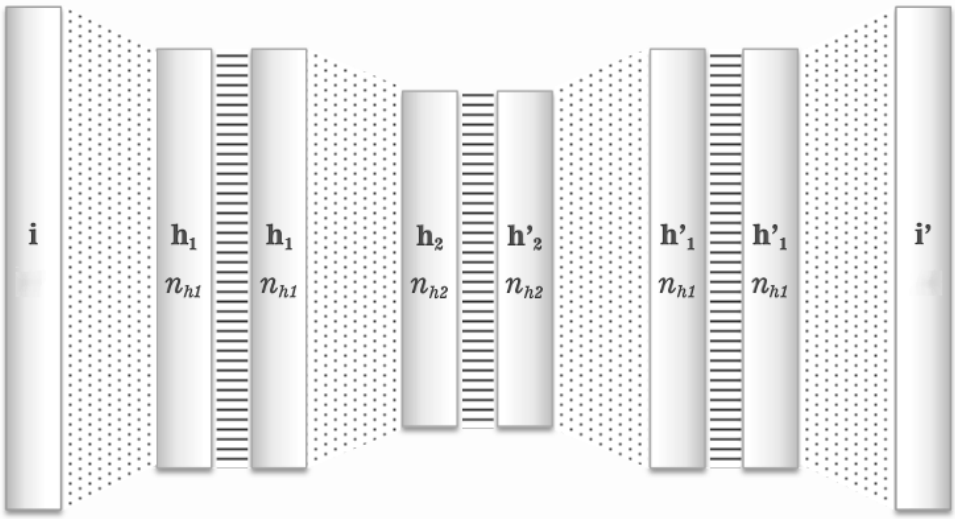

Figura 4.12: Autocodificador Profundo.

Etapa 6

En esta etapa se elimina la segunda parte del autocodificador profundo ($\mathbf{h}'_2 + \mathbf{h}'_1 + \mathbf{i}'$) (ver Figura 4.13) y se obtiene el codificador $\mathbf{i} + \mathbf{h}_1 + \mathbf{h}_2$; se pasan nuevamente los patrones utilizados en la etapa anterior y se generan las salidas que servirán de entrada para la siguiente etapa.

Etapa 7

Finalmente en esta última etapa se crea una red con el fin de inicializar los pesos de la capa de salida del clasificador profundo. Esta red consiste en n_{h2} neuronas en la capa de entrada y 6 en la capa de salida como se observa en la Figura 4.14). Para entrenar esta red

se utilizan los patrones generados en la etapa anterior y las salidas deseadas se obtuvieron a partir del corpus de datos, siendo una de las seis posibles emociones presentes.

Etapa 8

Una vez concluida la Etapa 7, ya se dispone del clasificador profundo pre-entrenado completamente. El entrenamiento en este caso actúa como un ajuste fino y supone encontrar mejores soluciones dada la inicialización de los pesos más cerca de mejores soluciones. En la Figura 4.15 se puede observar la arquitectura del mismo luego de completar todos los procesos anteriores.

4.2.3. Validación cruzada

Se utiliza la técnica estadística de validación cruzada para asegurar la generalización del clasificador y garantizar que los resultados obtenidos son independientes de la partición de entrenamiento y prueba.

En este trabajo, se utilizan 10 particiones de entrenamiento y prueba. En cada partición, se dividen los datos en conjuntos de entrenamiento y prueba, asignando un 80 % de los patrones aleatoriamente al conjunto de entrenamiento. Para evitar el sobreentrenamiento y detener el entrenamiento en el pico de máxima generalización, se subdividen el conjunto de entrenamiento en subconjuntos de estimación y validación.

Para cada época, se entrena (estima el modelo) con el conjunto de estimación y se prueba (valida el modelo) con el conjunto de validación. Se detiene el entrenamiento cuando se detecta el pico de generalización, es decir, cuando la tasa de errores en el conjunto de validación comienza a subir o se estanca. Se almacenan los pesos de épocas anteriores para poder utilizar los de mejor rendimiento en el conjunto de validación.

Dado que los datos están balanceados, es decir, se presentan misma cantidad de videos de cada una de las emociones, no es necesario un balanceo artificial para elegir las particiones.

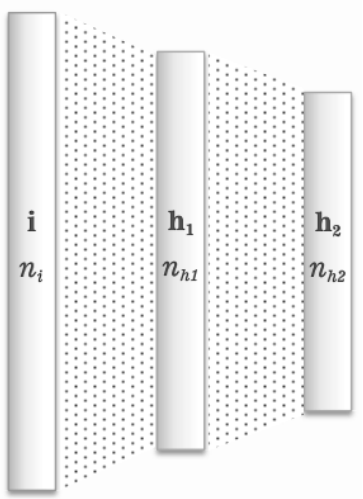

Figura 4.13: Red resultante de eliminar parte decodificadora del autocodificador profundo.

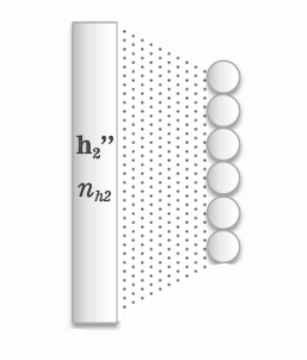

Figura 4.14: Red de la última capa.

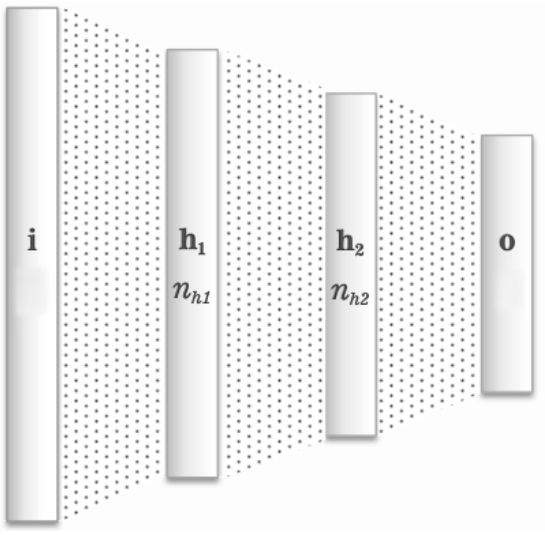

Figura 4.15: Clasificador profundo al final del proceso.

Capítulo 5

Experimentos y Resultados

Las pruebas fueron realizadas en una computadora de escritorio y una ultrabook con las siguientes características:

- **Desktop** Gigabyte M68M-S2P, Procesador AMD II X2 215 3200MHz, Memoria 4GB DDR II. Sistema Operativo Ubuntu Gnome 14.04.
- **Ultrabook** Asus UX31A, Procesador Intel Core i5 3800MHz, Memoria 4GB DDR III. Sistema Operativo Ubuntu Gnome 13.10.

Para la separación de la pista de audio de los archivos de video se utilizó el programa de línea de comandos `avconv` versión 9.11-6. Para la extracción de características de audio se utilizó MATLAB R2013a y la librería de procesamiento de voz VOICEBOX[1].

En la primera sección se especifican los criterios que se adoptaron a la hora de seleccionar la base de datos adecuada para las pruebas, y las características de la base de datos seleccionada.

Para la extracción de características de video se utilizó gcc versión 4.8.2, OpenCV 2.4.8 y librería de código abierto de detección de puntos faciales de interés flandmark 1.07[2].

Para la clasificación se utilizó el simulador de redes neuronales *Stuttgart Neural Networks Simulator* (SNNS), y se desarrollaron scripts en Bash (versión 4.3.8), Python (versión 2.7.6) y Perl 5 (versión 18, subversión 2).

5.1. Corpus de Datos

En esta sección se hace una revisión de las bases de datos disponibles en los diferentes centros de investigación que estudian el reconocimiento de emociones. Para lo cual, además

[1] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`
[2] `http://cmp.felk.cvut.cz/~uricamic/flandmark/`

de estudiar los trabajos de V. Bettapadura[11] y Z. Zeng [111] que resumen el estado del arte y las bases de datos de mayor renombre, se hizo una búsqueda global de las bases de datos multimodales creadas más recientemente.

5.1.1. Bases multimodales disponibles

Luego de haber realizado una exhaustiva búsqueda y considerando que no se dispone de presupuesto para adquirir bases de datos pagas, se logró obtener las siguientes:

- RML Emotion Database[3] [101]
  - origen canadiense;
  - 720 expresiones de emociones audiovisuales posadas;
  - 6 emociones humanas expresadas: enojo, asco, miedo, felicidad, tristeza y sorpresa; neutral (ausencia de emoción) al comienzo de cada frame de video;
  - entorno controlado, iluminado y fondo simple;
  - 10 sentencias de audio diferentes por cada emoción;
  - independiente de lenguaje y cultura (8 sujetos hablando 6 lenguajes diferentes: ingles, mandarín, urdú, punyabí, persa e italiano);
  - velocidad de muestreo: 22050 Hz, canal simple de 16-bit, 30 fps, formato AVI.
  - cantidad de trabajos que la utilizan: 11 [43, 51, 60, 93, 95, 102-104, 107, 108, 118].
- Surrey Audio-Visual Expressed Emotion (SAVEE) Database [4]
  - origen inglés;
  - 480 expresiones de emociones audiovisuales posadas;
  - 7 emociones humanas expresadas: enojo, asco, miedo, felicidad, tristeza, sorpresa y neutral;
  - entorno controlado, iluminado y fondo simple;
  - 15 sentencias de audio diferentes por cada emoción;
  - 60 marcadores para extracción de expresiones faciales;
  - lenguaje inglés;
  - velocidad de muestreo: 44100 Hz, 60 fps, formato AVI.
  - cantidad de trabajos que la utilizan: 3 [8, 10, 68].

[3] `http://www.rml.ryerson.ca/rml-emotion-database.html`
[4] `http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/Database.html`

- Belfast Naturalistic Database [5]
  - espontánea/inducida;
  - 298 clips obtenidos de programas televisivos y entrevistas;
  - etiquetado de emoción dimensional por 7 expertos;
  - lenguaje inglés;
  - video en formato MPEG, audio en formato .wav.
  - cantidad de trabajos que la utilizan: 22 [2, 13, 16, 17, 22, 24, 26, 27, 36, 37, 46, 48, 58, 65, 71, 73, 85, 86, 91, 92, 97, 100].

5.1.2. Criterios para la selección de la base de datos adecuada

En este apartado se definen los criterios de selección de la base de datos para realizar las pruebas de los métodos planteados en la sección anterior. Se tuvieron en cuenta los siguientes factores para determinar cual era la que mejor se adaptaba al problema a resolver:

1. etiquetado discreto de emociones;
2. suficiente cantidad de sujetos involucrados;
3. suficiente cantidad de muestras por emoción;
4. trabajos que utilicen la base de datos (cantidad y antigüedad);
5. calidad del videoclip;

El primer criterio es necesariamente una condición debido al método propuesto para la resolución del problema. Dado que por cada ejemplo de video se desea obtener la etiqueta de una emoción determinada se descartó la *Belfast Naturalistic Database* ya que ésta está etiquetada con un modelo dimensional, el cual refleja diferentes niveles de ciertas emociones presentes y no una en particular.

Los siguientes dos criterios fueron elegidos para dar al clasificador la capacidad de generalización; lo que se busca es que el sistema tenga la suficiencia de discernir entre emociones independientemente de la persona que está experimentado tal emoción. Es entonces preferible contar con una base de datos que contenga una variedad de actores involucrados. Se consideró, luego de estudiar el estado del arte, que al menos 5 actores y 400 muestras (balanceadas) eran suficientes para lograr tal cometido.

El cuarto criterio es de relevancia porque es un indicador del uso o difusión de cada base de datos en la comunidad científica. La cantidad y calidad de trabajos que hayan

[5] `http://belfast-naturalistic-db.sspnet.eu/`

optado por una base de datos es entonces importante a la hora de realizar la elección. Es además relevante si los trabajos que citan la base de datos son recientes, o si por el contrario la base de datos ya no es utilizada por la comunidad científica.

El quinto criterio fue elegido para lograr una correcta evaluación de emociones de forma "instantánea", por lo que se precisó que la base de datos haya sido creada con cámaras de grabación a 30 fps.

En función de estos criterios se decidió optar por la *RML Emotion Database* dado que, además de cumplir con los criterios planteados de manera satisfactoria (base de datos de emociones prototípicas con frecuencia de muestreo de 30 fps, de suficiente cantidad de actores y muestras, citada con frecuencia en trabajos actuales de la comunidad científica), los actores no presentan ningún tipo de marcador facial lo que hace a esta base de datos conveniente para probar el sistema planteado.

5.2. Extracción de características en audio

Para la extracción de características se desarrollaron los siguientes scripts de MATLAB:

- **F0_energy.m**: estima media y desvío estándar de la frecuencia fundamental y de la energía de corta duración de una señal de audio.
- **MLS.m**: estima 30 coeficientes MLS.
- **features.m**: recibe un archivo de audio, por ejemplo `an10.wav`, calcula 12 medias de los coeficientes MFCC, las medias de sus derivadas primera y segunda, llama a los dos scripts anteriores para obtener las demás características, concatena todas las características y guarda un archivo `an10-audio_features.csv`.
- **get_audio_features.m**: corre `features.m` para todos los archivos de audio de una carpeta.

Se desarrolló un script en bash, **extract_audio_features.sh**, que corre este último script para todas las carpetas de la base de datos. Por lo tanto, para cada archivo de video `*.avi` de la base de datos se cuenta con un archivo `*-audio_features.csv`.

Para la detección de *frames* con voz se utilizó el detector de actividad de voz VADSOHN de la librería VOICEBOX. Se optimizaron los siguientes parámetros:

- **pp.pr**: umbral de probabilidad de discurso. Se fijó en 0.99.
- **pp.ts**: media de ráfaga de voz en milisegundos. Se fijó en 800.
- **pp.tn**: media de largo de silencios en milisegundos. Se fijó en 15.
- **pp.ta**: constante para suavizar la SNR estimada. Se fijó en 0.9

Con estos parámetros se obtuvieron los mejores resultados en la detección del segmento de la señal de audio en los que el hablante pronuncia la frase con emoción. Esta detección se utilizó en la detección de la frecuencia fundamental, de la energía de corta duración y en la detección del primer *frame* de video que se corresponde al comienzo de la pronunciación de la frase. Esta detección será detallada más adelante, en la extracción de características en video.

La estimación de la frecuencia fundamental fue realizada utilizando la implementación del método PEFAC de la librería VOICEBOX. Se realizó con ventanas de 10 milisegundos.

La energía de corta duración fue estimada utilizando ventanas de Hamming de 10 milisegundos.

Las medias de los 12 MFCCs junto con las medias de sus derivadas fueron calculadas utilizando la función `melcepst` de la librería VOICEBOX, con los siguientes parámetros:

- Tipo de ventana: ventana de Hamming.
- Tamaño de ventana: 23 milisegundos.
- Desplazamiento: 11.5 milisegundos

Se desarrolló un script en bash **concatenate_audios.sh** para concatenar todos los archivos de características en un archivo `all_audio_features.csv`. Luego se desarrolló un script en matlab para normalizar las características al rango $[-1, 1]$. Además se utilizó el script **generate_partitions.py** para generar las particiones necesarias para la validación cruzada.

5.3. Extracción de características en video

La extracción de características en video se realizó mediante el desarrollo de los siguientes programas y scripts:

- Script de matlab **voice_t0.m** que obtiene el *frame* de video en el que se comienza a pronunciar la frase y el *frame* en el que termina.
- Script de matlab **get_voiced_frames.m** que corre el script anterior para todos los archivos de audio y genera un archivo `*-voiced_frames.txt` para cada archivo de la base de datos.
- Script en bash **prom_voiced_frames.sh** que obtiene el promedio de frames con voz para todos los videos a partir de los archivos generados en el script anterior.
- Programa en C++ **VideoFeaturesExtraction** que utiliza la librería de procesamiento de imágenes OpenCv y la librería flandmark.

- Script en bash **extract_video_features.sh** que corre el programa anterior para todos los archivos de la base de datos y genera un archivo `*-video_features.csv` para cada archivo de la base de datos.
- Script en bash **concatenate_videos.sh** que agrupa todos los archivos de características en un archivo llamado `all_video_features.csv`.
- Script en python **generate_partitions.py** que genera las particiones necesarias para la validación cruzada.
- Programa en C++ **PCA** que aplica PCA sobre las características de video. Este programa se corre sobre cada una de las particiones, genera los espacios a partir del conjunto de datos de entrenamiento, y luego proyecta todos los vectores de características para obtener las representaciones en los nuevos espacios.

En el programa **VideoFeaturesExtraction** se optimizaron los siguientes parámetros:

- Tamaño de las regiones de interés: existe el compromiso entre mayor detalle o menores datos para realizar PCA. Se fijó en 70×35 píxeles para ambas regiones de interés.
- Número de componentes principales a retener al aplicar PCA. Se eligió 70 para obtener el mismo número de características que de audio y que el clasificador tenga información equilibrada de ambos medios.

Dadas las características de la base de datos elegida, hubo que manejar excepciones en casos en que la librería flandmark no detectaba los puntos de interés. Esto pasó en casos en que el sujeto rotaba la cabeza en alguna dirección y por lo tanto la cara no aparece en la imagen, como se puede ver en la Figura 5.1. Debido a que esto solo sucedió en 3 videos en un reducido número de *frames*, se decidió descartar los mismos y repetir la información de los anteriores.

Figura 5.1: Detección de los puntos de interés a pesar de la rotación de la cara (izquierda) y error en la detección de los puntos de interés debido a demasiada rotación (derecha).

5.4. Resultados y discusión

En la etapa de clasificación se desarrollaron los siguientes scripts en Python con el fin de preparar los datos para ser procesados por el simulador de redes neuronales SNNS:

- **generate_SNNS_data.py**: adapta las particiones para generar los patrones de entrada para el primer autocodificador (entrada=salida) en el formato necesario para el uso con SNNS.
- **generate_SNNS_data_class.py**: adapta las particiones para generar los patrones de entrada al clasificador final (entradas y salidas deseadas) en el formato necesario para el uso con SNNS.

Se adaptaron una serie de scripts en Perl provistos por los Directores del Proyecto y desarrollados por Neri Cibau, y con los que se obtuvieron los resultados publicados en [21]. Estos scripts fueron adaptados para realizar la construcción de los diferentes autocodificadores y clasificadores necesarios usando el simulador SNNS.

Las pruebas de clasificación se diseñaron para probar el rendimiento de los vectores de características de audio y video, y con los vectores de características fijos se probaron diferentes arquitecturas de la red neuronal profunda (como fue detallado en la Sección 4.2.2).

Se probó con el vector de características de audio de 46 coeficientes definido anteriormente, $\mathbf{fva}_{46}$, y con otro denominado $\mathbf{fva}_{70}$, en el que además se utilizan las primeras y segundas derivadas de los MFCC.

En cuanto a las características de video, se probó con el vector de 70 coeficientes definido anteriormente, $\mathbf{fvv}_{70}$, y con un vector que denominamos $\mathbf{fvv}_{85}$ que asegura retener al menos el 90 % de la varianza por conjunto de emociones al aplicar PCA.

De esta manera, se definen los siguientes vectores de características para las pruebas, de acuerdo a las combinaciones posibles, como se observa en la Tabla 5.1.

Tabla 5.1: Vectores de características definidos para las pruebas

N^o	Nombre	Combinación
1	$\mathbf{fv}_{140}$	$\mathbf{fva}_{70} + \mathbf{fvv}_{70}$
2	$\mathbf{fv}_{155}$	$\mathbf{fva}_{70} + \mathbf{fvv}_{85}$
3	$\mathbf{fv}_{116}$	$\mathbf{fva}_{46} + \mathbf{fvv}_{70}$
4	$\mathbf{fv}_{131}$	$\mathbf{fva}_{46} + \mathbf{fvv}_{85}$

Para cada uno de los vectores de características se definieron diferentes arquitecturas neuronales, como se detalló en la Sección 4.2.2.

Para cada una de estas estructuras se variaron las funciones de activación de cada una de las capas. Se utilizaron la *tangente hiperbólica*, la *logística*, la *identidad* y la *senoidal*

con los parámetros de entrenamiento estándar y se comprobó que la función senoidal proporciona los mejores resultados de clasificación y de autocodificación para los datos de prueba en cada una de las capas. Es por ésto que para la optimización de los parámetros de entrenamiento de ésta y de las demás arquitecturas se utilizó entonces la función de activación senoidal para cada una de las capas.

El entrenamiento de cada una de las etapas se realizó utilizando el algoritmo de *Retropropagación con término de momento* implementado en SNNS. Se optimizaron los siguientes parámetros de entrenamiento:

- La tasa de aprendizaje μ (valores típicos: 0,1 ... 1,0).
- El término de momento η (valores típicos: 0,0 ... 1,0).
- Parámetro para evitar caer en zonas planas del espacio de solución c (valores típicos: 0,0 ... 0,25).
- El error máximo tolerado en la salida de la i-ésima neurona de la capa de salida, e_{imax} (valores típicos: 0 ... 0,1).

Se utiliza la regla de clasificación *winner-takes-all*, en la que la salida con mayor nivel de activación determina la clase de salida de la red. Para cada arquitectura se optimizaron los parámetros de entrenamiento y las épocas y se llegó a un conjunto de parámetros que se muestra en la Tabla 5.2.

Tabla 5.2: Parámetros de entrenamiento optimizados.

Etapa de entrenamiento	Épocas	μ	η	c	e_{imax}
Etapa 1	3000	0.1	0.5	0.0	0.07
Etapa 3	2500	0.1	0.5	0.0	0.07
Etapa 5	4500	0.02	0.03	0.07	0.07
Etapa 7	5000	0.02	0.01	0.0	0.0
Etapa 8	5000	0.01	0.02	0.07	0.07

Para comparar el rendimiento del método propuesto con cada vector de características, se implementó un clasificador estándar del tipo MLP. En la Tabla 5.3 se pueden observar los parámetros de entrenamiento para los clasificadores MLP implementados para las pruebas. Se utilizaron funciones de activación logísticas para todas las capas de los clasificadores MLP.

Tabla 5.3: Parámetros de entrenamiento optimizados para MLP.

	Épocas	μ	η	c	e_{imax}
MLP	3000	0.01	0.02	0.07	0.07

Se realizaron pruebas con diferentes arquitecturas de clasificador estándar MLP de 3 capas. Para asegurar la correcta comparación de los clasificadores, se tuvo en cuenta

que ambos tengan la misma cantidad de pesos o *parámetros libres*. Por lo tanto, en la capa oculta de cada MLP se seleccionaron los números de neuronas correspondientes para asegurar esto. Sin embargo, también se realizaron pruebas con diferentes cantidades de neuronas en la capa oculta, que en algunos casos produjeron mejores resultados.

5.4.1. Experimentos multimodales

Se diseñaron 5 arquitecturas para cada uno de los vectores de características $\mathbf{fv}_{140}$, $\mathbf{fv}_{155}$, $\mathbf{fv}_{116}$ y $\mathbf{fv}_{131}$. Se define la siguiente notación para denominar las arquitecturas propuestas: se denomina CPi_X a la arquitectura número i para el vector de características $\mathbf{fv}_X$.

En las Tablas 5.4 a 5.7 se pueden apreciar los errores cuadráticos medios en las etapas de entrenamiento 1, 3 y 5, el porcentaje de desaciertos de la etapa 7 y el porcentaje de clasificación para las diferentes arquitecturas definidas en cada uno de estos vectores de características. Además, se hace la comparación con el clasificador estándar MLP para las arquitecturas correspondientes.

En la etapa 7 se inicializan los pesos que unen la última capa oculta y la capa de salida, es decir se entrenan las redes de la forma $\mathbf{h_2} + \mathbf{o}$. Esta red ya no es un autocodificador, sino un clasificador, pero no presenta capa oculta y sólo se construye para inicializar los pesos que unen la última capa del clasificador $\mathbf{i} + \mathbf{h_1} + \mathbf{h_2} + \mathbf{o}$. Es esperable, como se puede ver en los experimentos, que no se obtengan buenos resultados de clasificación con esta red, debido a que es una red neuronal sin capa oculta, por lo que no cuenta con capacidad para separar clases no lineales. Sin embargo, es una mejor aproximación a una inicialización aleatoria de los pesos para la siguiente etapa.

En las Figuras 5.2 a 5.5 se pueden apreciar las variaciones de los errores de clasificación a través de las épocas para las mejores particiones de las arquitecturas ganadoras para un vector de características. En cada imagen se compara, dado un vector de características, la arquitectura ganadora del método propuesto con el clasificador estándar MLP.

Tabla 5.4: Errores MSE y % de clasificación multimodal para $\mathbf{fv}_{140}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	**$\mathbf{h_1}$**	**$\mathbf{h_2}$**	**o**					
$CP1_{140}$	**140**	**90**	**25**	**6**	**0,41**	**0,91**	**1,36**	**60,42**	**75,694**
$CP2_{140}$	140	90	50	6	0,41	0,44	1,15	59,58	72,777
$CP3_{140}$	140	100	45	6	0,28	0,67	1,02	57,50	73.333
$CP4_{140}$	140	100	50	6	0,28	0,60	0,95	55,83	74,22
$\mathbf{CP5}_{140}$	140	100	60	6	0,28	0,47	0,85	56,94	74,306
$\mathbf{MLP}_{140}$	**140**	**103**		**6**	**-**	**-**	**-**	**-**	**73,611**

En la Tabla 5.4 se puede observar que, para el vector de características $\mathbf{fv}_{140}$, el clasificador profundo implementado en este trabajo tuvo mejor rendimiento que el clasificador estándar MLP por un 2 %. En este caso, el mejor resultado para el clasificador estándar

MLP se obtuvo con la cantidad de neuronas elegida para equiparar la cantidad de pesos con la arquitectura ganadora, es decir con la arquitectura $\mathbf{CP5}_{140}$.

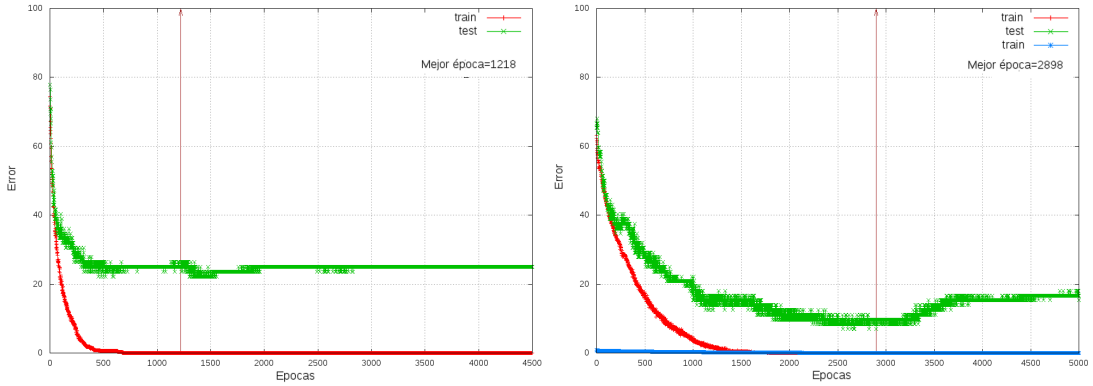

Figura 5.2: Errores de clasificación para MLP_{140}(izquierda) y para CP1_{140} (derecha).
Rojo: error de clasificación de entrenamiento. Verde: error de clasificación de validación. Azul: MSE de entrenamiento.

En la Tabla 5.5 se puede ver que el vector de características $\mathbf{fv}_{155}$ resultó en menores rendimientos de clasificación que con el vector $\mathbf{fv}_{140}$, tanto para el método propuesto como para el clasificador estándar. El método propuesto mejoró el clasificador estándar por un 2 %, utilizando una arquitectura con menor número de neuronas. En la Figura 5.3 se puede apreciar el error de clasificación de las mejores particiones para la arquitectura CP5_{155} y para el clasificador estándar MLP_{155}.

Tabla 5.5: Errores MSE y % de clasificación multimodal para $\mathbf{fv}_{155}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	$\mathbf{h_1}$	$\mathbf{h_2}$	**o**					
CP1_{155}	155	70	40	6	0,76	0,25	1,19	61,81	71,945
CP2_{155}	155	75	35	6	0,71	0,39	1,27	61,81	71,389
CP3_{155}	**155**	**80**	**50**	**6**	**0,67**	**0,22**	**1,05**	**58,89**	**72,777**
CP4_{155}	155	90	50	6	0,56	0,33	1,03	58,05	72,36
CP5_{155}	155	100	50	6	0,44	0,46	1,36	59,58	70,556
MLP_{155}	**155**	**75**		**6**	**-**	**-**	**-**	**-**	**70,972**
MLP_{155}	155	104		6	-	-	-	-	67,917

La Tabla 5.6 muestra los errores de clasificación para el vector de características $\mathbf{fv}_{116}$. Con este vector de características se obtuvo un rendimiento de casi un 80 % de clasificación para la mejor de las arquitecturas, lo cual mejora al clasificador estándar en un 3 %. Este vector de características consiste en 46 características de audio y 70 características de video. En la Figura 5.4 se ven los errores de clasificación por época de las mejores particiones de la arquitectura CP4_{116} y para el clasificador estándar MLP_{116}.

Por último, en la Tabla 5.7 se observa que para el vector de características $\mathbf{fv}_{131}$ se obtuvo un rendimiento del 78,3 %, lo que es aproximadamente un 0,5 % mejor que el rendimiento del clasificador estándar para el mismo vector de características.

Se observa que los mejores resultados se obtuvieron con los vectores de características $\mathbf{fv}_{116}$ y $\mathbf{fv}_{131}$, que son los vectores de características que utilizan el vector de característi-

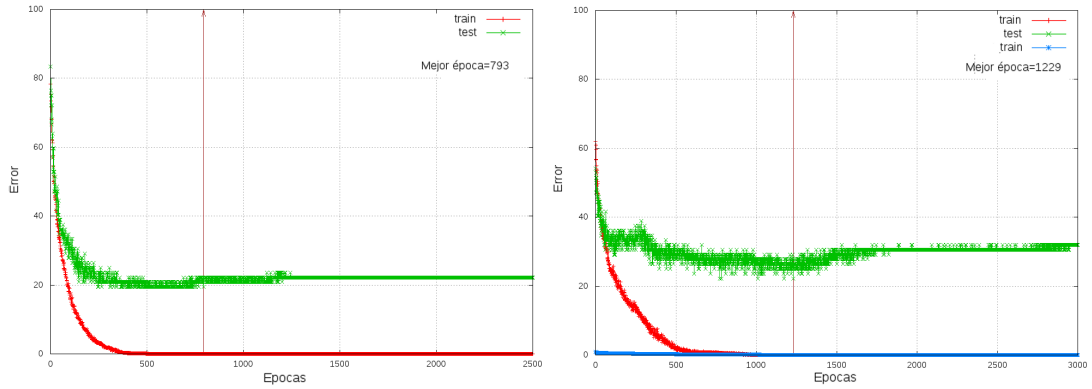

Figura 5.3: Errores de clasificación para MLP_{155}(izquierda) y para $CP3_{155}$ (derecha).
Rojo: error de clasificación de entrenamiento. Verde: error de clasificación de validación. Azul: MSE de entrenamiento.

cas de audio $\mathbf{fva}_{46}$. De estos resultados se puede concluir que agregar las medias de las derivadas de los MFCC empeora los rendimientos de clasificación multimodal.

Tabla 5.6: Errores MSE y % de clasificación multimodal para $\mathbf{fv}_{116}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	$\mathbf{h_1}$	$\mathbf{h_2}$	**o**					
$CP1_{116}$	116	70	35	6	0,36	0,24	0,72	61,39	77,916
$CP2_{116}$	**116**	**100**	**40**	**6**	**0,09**	**0,53**	**0,65**	**59,17**	**79,724**
$CP3_{116}$	116	100	50	6	0,09	0,41	0,55	57,36	78,333
$CP4_{116}$	116	100	60	6	0,09	0,32	0,47	57,36	78,661
$CP5_{116}$	116	105	60	6	0,08	0,31	0,47	57,08	77,638
$\mathbf{MLP}_{116}$	**116**	**50**		**6**	**-**	**-**	**-**	**-**	**76,807**
MLP_{116}	116	138		6	-	-	-	-	76,805

Tabla 5.7: Errores MSE y % de clasificación multimodal para $\mathbf{fv}_{131}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	$\mathbf{h_1}$	$\mathbf{h_2}$	**o**					
$CP1_{131}$	131	80	45	6	0,36	0,18	0,70	56,53	76,667
$CP2_{131}$	131	85	40	6	0,32	0,27	0,72	58,75	74,999
$\mathbf{CP3}_{131}$	**131**	**90**	**35**	**6**	**0,27**	**0,40**	**0,76**	**58,20**	**78,334**
$CP4_{131}$	131	95	35	6	0,22	0,47	0,76	56,67	77,5
$CP5_{131}$	131	100	45	6	0,17	0,43	0,65	56,53	77,916
$\mathbf{MLP}_{131}$	**131**	**66**		**6**	**-**	**-**	**-**	**-**	**77.78**
MLP_{131}	131	110		6	-	-	-	-	76.39

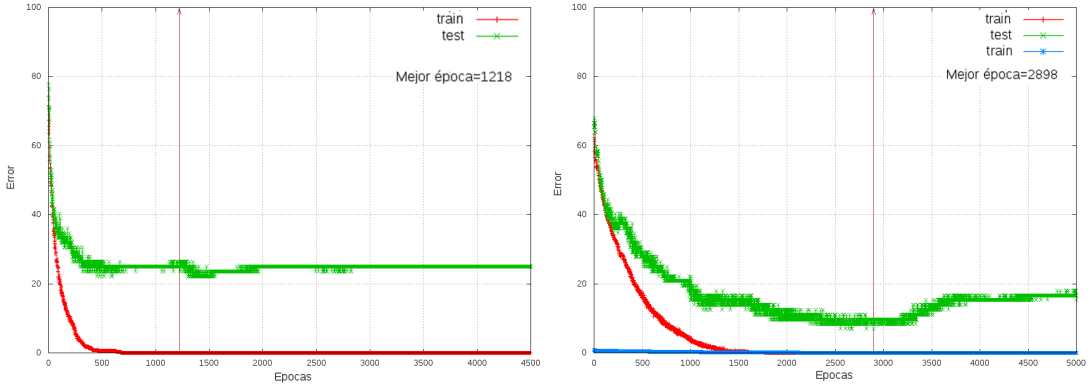

Figura 5.4: Errores de clasificación para MLP_{116}(izquierda) y para $CP2_{116}$ (derecha).
Rojo: error de clasificación de entrenamiento. Verde: error de clasificación de validación. Azul: MSE de entrenamiento.

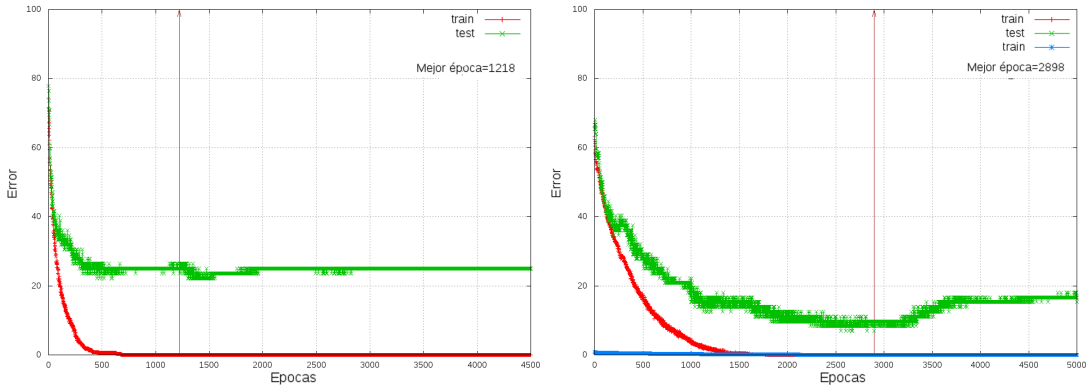

Figura 5.5: Errores de clasificación para MLP_{131}(izquierda) y para $CP3_{131}$ (derecha).
Rojo: error de clasificación de entrenamiento. Verde: error de clasificación de validación. Azul: MSE de entrenamiento.

5.4.2. Experimentos monomodales

Se realizaron experimentos monomodales para evaluar la contribución de las características de audio y de video por separado en los resultados de clasificación. Se utilizaron para esto los 4 vectores de características definidos anteriormente: $\mathbf{fva}_{46}$, $\mathbf{fva}_{70}$, $\mathbf{fvv}_{70}$ y $\mathbf{fvv}_{85}$. Se definieron 4 arquitecturas de clasificador profundo para cada uno de estos vectores de características, utilizando las funciones de activación y los parámetros de entrenamiento optimizados para el clasificador multimodal. Siguiendo la misma notación, denominamos CPAi_X a la arquitectura número i para el vector de características $\mathbf{fva}_X$, y CPVi_X a la arquitectura número i para el vector de características $\mathbf{fvv}_X$. Además se probó con el clasificador estándar MLP con los parámetros de entrenamiento definidos anteriormente para comparar los rendimientos de clasificación.

En las Tablas 5.8 y 5.9 se pueden apreciar los errores cuadráticos medios en las etapas de entrenamiento 1, 3 y 5, el porcentaje de desaciertos de la etapa 7 y el porcentaje de clasificación de las diferentes arquitecturas definidas para los vectores de características $\mathbf{fva}_{46}$ y $\mathbf{fva}_{70}$. Además, se hace la comparación con el clasificador estándar MLP para las arquitecturas correspondientes.

Tabla 5.8: Errores MSE y % de clasificación monomodal para $\mathbf{fva}_{46}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	**h_1**	**h_2**	**o**					
$CPA1_{46}$	46	25	13	6	0,06	0,13	0,18	56,11	75,694
$CPA2_{46}$	46	30	15	6	0,03	0,12	0,15	54,86	73,749
$CPA3_{46}$	46	35	15	6	0,02	0,13	0,15	53,05	75,416
$CPA4_{46}$	**46**	**40**	**20**	**6**	**0,02**	**0,06**	**0,10**	**49,86**	**77,084**
$MLPA_{46}$	**46**	**53**		**6**	**-**	**-**	**-**	**-**	**76,945**

Tabla 5.9: Errores MSE y % de clasificación monomodal para $\mathbf{fva}_{70}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	**h_1**	**h_2**	**o**					
$CPA1_{70}$	70	35	15	6	0,18	0,38	0,56	61,11	71,527
$CPA2_{70}$	70	40	20	6	0,13	0,29	0,42	60,55	70,973
$CPA3_{70}$	70	50	25	6	0,06	0,25	0,33	57,36	70,138
$CPA4_{70}$	**70**	**60**	**30**	**6**	**0,05**	**0,17**	**0,25**	**50,70**	**71,944**
$MLPA_{70}$	**70**	**81**		**6**	**-**	**-**	**-**	**-**	**69,305**

Se puede observar que en los experimentos monomodales de audio se mantienen los resultados dados en los experimentos multimodales. Se obtienen mejores rendimientos de clasificación con el vector de características $\mathbf{fva}_{46}$, con el que se llega a un 77 % de aciertos, resultado que es levemente mejor que el obtenido con el clasificador estándar MLP. Con el vector $\mathbf{fva}_{70}$, en cambio, se obtuvo un rendimiento del 72 % de aciertos aproximadamente.

De manera análoga, en las Tablas 5.10 y 5.11 se pueden apreciar los errores cuadráticos medios en las etapas de entrenamiento 1, 3 y 5, el porcentaje de desaciertos de la etapa 7 y el porcentaje de clasificación de las diferentes arquitecturas definidas para los vectores de

características $\mathbf{fvv}_{70}$ y $\mathbf{fvv}_{85}$. Además, se hace la comparación con el clasificador estándar MLP para las arquitecturas correspondientes.

Tabla 5.10: Errores MSE y % de clasificación monomodal para $\mathbf{fvv}_{70}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	$\mathbf{h_1}$	$\mathbf{h_2}$	**o**					
$CPV1_{70}$	70	35	15	6	0,19	0,16	0,47	70,97	72,36
$CPV2_{70}$	**70**	**45**	**20**	**6**	**0,12**	**0,16**	**0,37**	**72,50**	**73,473**
$CPV3_{70}$	70	50	30	6	0,09	0,09	0,25	69,17	71,945
$CPV4_{70}$	70	55	25	6	0,06	0,17	0,30	72,22	71,389
$MLPV_{70}$	**70**	**55**		**6**	**-**	**-**	**-**	**-**	**71,389**

Tabla 5.11: Errores MSE y % de clasificación monomodal para $\mathbf{fvv}_{85}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	$\mathbf{h_1}$	$\mathbf{h_2}$	**o**					
$CPV1_{85}$	85	35	15	6	0,23	0,15	0,52	72,78	74,167
$CPV2_{85}$	**85**	**40**	**20**	**6**	**0,20**	**0,11**	**0,43**	**73,06**	**75,416**
$CPV3_{85}$	85	45	25	6	0,17	0,10	0,36	69,31	74,307
$CPV4_{85}$	85	50	35	6	0,14	0,05	0,26	65,27	72,217
$MLPV_{70}$	**70**	**55**		**6**	**-**	**-**	**-**	**-**	**71,804**

Puede observarse en los experimentos monomodales de video que se obtiene un mejor rendimiento cuando se retiene un mayor porcentaje de la varianza, como era esperable. Para ambos vectores de características, el método propuesto obtuvo mejores resultados que el clasificador estándar.

Otro análisis interesante es la comparación entre la clasficación monomodal y multimodal. Si comparamos los rendimientos obtenidos con los vectores de características $\mathbf{fva}_{46}$ y $\mathbf{fvv}_{85}$ con los rendimientos obtenidos en la clasificación multimodal con el vector $\mathbf{fv}_{131}$, que es la concatenación de estos dos vectores, se puede observar que el enfoque multimodal mejora los rendimientos de clasificación. Con el vector $\mathbf{fva}_{46}$ el mejor resultado obtenido fue de un 77,08 %, y con el vector $\mathbf{fvv}_{85}$ se obtuvo un 75,41 %, mientras que con el vector $\mathbf{fv}_{131}$ se obtuvo un 78,33 % de aciertos.

5.4.3. Experimentos multimodales monoidioma

Dado que la base de datos *RML Emotion Database* cuenta con emociones expresadas en 6 idiomas distintos con diferenes variantes culturales, el clasificador de emociones diseñado distingue emociones independientemente del idioma y de la variante cultural. Para probar el método propuesto frente a un único idioma, se decidió realizar pruebas de clasificación de emociones sólo en idioma inglés.

Para obtener el vector de características de video, se realiza el mismo procedimiento detallado en la Sección 4.1 pero esta vez sobre un conjunto reducido de la base de datos; el mismo posee 30 videos por emoción con personajes hablando sólo en idioma inglés. De

esta manera la muestra queda balanceada. El vector obtenido $\mathbf{fvve}_{24}$ tiene 24 coeficientes, lo que implica retener el 100 % de la varianza al aplicar PCA.

En cuanto a las características de audio, se probaron, nuevamente, las 2 alternativas: un vector $\mathbf{fvae}_{70}$ manteniendo las medias de las primeras y segundas derivadas de los MFCC, y otro vector $\mathbf{fvae}_{46}$, sin éstas.

De esta manera, se definen los siguientes vectores de características para las pruebas en idioma inglés, de acuerdo a las combinaciones posibles, como se observa en la Tabla 5.12.

Tabla 5.12: Vectores de características definidos para las pruebas en idioma inglés.

N^o	Nombre	Combinación
1	$\mathbf{fve}_{70}$	$\mathbf{fvae}_{46}$ + $\mathbf{fvve}_{24}$
2	$\mathbf{fve}_{94}$	$\mathbf{fvae}_{70}$ + $\mathbf{fvve}_{24}$

Se definieron 5 arquitecturas de clasificador profundo para cada uno de estos vectores de características, utilizando las funciones de activación y los parámetros de entrenamiento optimizados para el clasificador para todas las emociones. Siguiendo la misma notación, denominamos CPEi_X a la arquitectura número i para el vector de características $\mathbf{fve}_X$.

En las Tablas 5.13 y 5.14 se pueden apreciar los errores cuadráticos medios en las distintas etapas de entrenamiento y el porcentaje de clasificación para las diferentes arquitecturas definidas en cada uno de estos vectores de características. Además, se hace la comparación con el clasificador estándar MLP para las arquitecturas correspondientes.

Tabla 5.13: Errores MSE y % de clasificación monoidioma para $\mathbf{fve}_{70}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	$\mathbf{h_1}$	$\mathbf{h_2}$	**o**					
CPE1_{70}	70	35	15	6	0,31	0,42	0,90	28,89	77,222
CPE2_{70}	70	45	25	6	0,21	0,29	0,60	28,89	76,666
CPE3_{70}	70	50	20	6	0,19	0,43	0,71	30,00	75,556
CPE4_{70}	**70**	**50**	**25**	**6**	**0,19**	**0,30**	**0,60**	**33,33**	**78,332**
CPE5_{70}	70	60	40	6	0,17	0,10	0,30	29,44	77,778
MLP_{70}	**70**	**60**		**6**	**-**	**-**	**-**	**-**	**74,445**

En la Tabla 5.13 se puede ver que los resultados de clasificación son congruentes con los obtenidos en la clasificación multilenguaje. Se puede observar que se logra una tasa promedio de clasificación del 78 % lo cual era esperable dado que el clasificador multilenguaje entrenado con el vector de características $\mathbf{fve}_{116}$ y $\mathbf{fve}_{131}$ (ambos con 46 características de audio) logra una clasificación promedio del 79 % aproximadamente. Esto permite concluir que el método propuesto de extracción de caracterísitcas en audio y video es invariante al idioma y a la cultura de las personas.

En la Tabla 5.14 puede verse una vez más que, como en la clasificación multilenguaje, añadir las medias de las primeras y segundas derivadas de los MFCC no resultó una mejoría en la clasificación. Por el contrario, los resultados empeoran notablemente. Esto

Tabla 5.14: Errores MSE y % de clasificación monoidioma para $\mathbf{fve}_{100}$

Arquitectura	Estructura				MSE Etapa 1	MSE Etapa 3	MSE Etapa 5	% Desaciertos Etapa 7	% Clasificación
	i	**h_1**	**h_2**	**o**					
$CPE1_{94}$	94	50	25	6	0,63	0,90	1,82	34,44	67,778
$CPE2_{94}$	94	60	25	6	0,51	1,05	1,79	27,78	63,333
$CPE3_{94}$	94	65	30	6	0,49	0,81	1,52	36,67	68,889
$CPE4_{94}$	**94**	**70**	**30**	**6**	**0,48**	**0,83**	**1,52**	**30,00**	**70,001**
$CPE5_{94}$	94	80	40	6	0,46	0,40	1,07	33,89	68,89
MLP_{94}	**94**	**89**		**6**	**-**	**-**	**-**	**-**	**69,445**

puede deberse a que se añade información que no es tan significativa para la clasificación a costa de añadir neuronas y complejidad al clasificador.

5.4.4. Discusiones finales

En la Tabla 5.15 se puede observar un resumen de los resultados multimodales para los distintos vectores de características que se evaluaron en este trabajo. Se puede observar que en todos los casos el método propuesto fue superior en rendimiento de clasificación con respecto al método estándar. El mejor resultado de clasificación con el método propuesto fue de un 79,724 %, y se obtuvo con el vector de características $\mathbf{fv}_{116}$. Por otro lado, el mejor resultado obtenido con clasificador estándar MLP fue de un 77,78 % de aciertos con el vector de características $\mathbf{fv}_{131}$.

Tabla 5.15: Resumen de resultados multimodales.

Vector de características	Método propuesto	Método estándar
$\mathbf{fv}_{140}$	75,694 %	73,611 %
$\mathbf{fv}_{155}$	72,777 %	70,972 %
$\mathbf{fv}_{116}$	**79,724 %**	76,807 %
$\mathbf{fv}_{131}$	78,334 %	77,78 %

Además, se puede observar que los mejores resultados se obtuvieron con los vectores de características que utilizan el vector de características de audio $\mathbf{fva}_{46}$, es decir, cuando no se utilizan las medias de las primeras y segundas derivadas de los MFCC.

En la Tabla 5.16 se pueden observar los resultados monomodales para los distintos vectores de características. Se puede apreciar que el método propuesto fue superior en todos los casos, y que los mejores resultados se obtienen utilizando el vector de características de audio $\mathbf{fva}_{46}$ y el vector de características de video $\mathbf{fvv}_{85}$, con los cuales se llega a un 77,084 % y un 75,416 % respectivamente. Estos vectores de características combinados forman el vector $\mathbf{fv}_{131}$, con el que se obtuvo un 78,334 % en el enfoque multimodal, por lo que se puede observar que el enfoque multimodal mejora los rendimientos de clasificación monomodales. Por otro lado, con los vectores $\mathbf{fva}_{46}$ y $\mathbf{fva}_{70}$ se obtiene el vector de

características $\mathbf{fv}_{116}$, con el cual se llegó a un 79,724 % en el enfoque multimodal, lo que es casi un 3 % mejor que el mejor resultado obtenido con estos vectores individualmente.

Tabla 5.16: Resumen de resultados monomodales.

Vector de características	Método propuesto	Método estándar
$\mathbf{fva}_{46}$	**77,084 %**	76,945 %
$\mathbf{fva}_{70}$	71,944 %	69,305 %
$\mathbf{fvv}_{70}$	73,473 %	71,389 %
$\mathbf{fvv}_{85}$	**75,416 %**	71,804 %

En la Tabla 5.17 se pueden observar los resultados multimodales monoidioma para el método propuesto y el método estándar. Se observa que nuevamente es superior el rendimiento cuando no se utilizan las medias de las primeras y segundas derivadas de los MFCC. Además, se llega a resultados similares que en la clasificación multiidioma, con lo que se puede concluir que los resultados de este trabajo son invariantes al idioma y a la cultura de las personas.

Tabla 5.17: Resumen de resultados multimodales monoidioma.

Vector de características	Método propuesto	Método estándar
$\mathbf{fve}_{70}$	**78,332 %**	74,445 %
$\mathbf{fve}_{94}$	70,001 %	69,445 %

El mejor resultado para el método propuesto se obtuvo con la arquitectura $\mathbf{CP2}_{116}$, es decir 116 neuronas en la capa de entrada, 100 neuronas en la primer capa oculta, 50 neuronas en la segunda capa oculta y 6 neuronas en la capa de salida, utilizando el vector de características $\mathbf{fv}_{116}$. En la Tabla 5.18 se puede ver la matriz de confusión promedio para esta red.

En la matriz de confusión se puede observar que la emoción más difícil de distinguir fue asco, para la cual se obtuvo una tasa de reconocimiento del 70 %. Esta emoción es confundida principalmente con miedo, en el 10 % de los casos. Por otro lado, la emoción más distinguible fue ira, con una porcentaje de acierto del 88,33 %. Así mismo, el resto de las emociones tienen unas tasas de acierto que van del 78 % al 82 %.

Tabla 5.18: Matriz de confusión promedio de las 10 particiones de la arquitectura $CP1_{116}$.

Clase	Ira	Asco	Miedo	Alegría	Tristeza	Sorpresa
Ira	**10,6**	0,2	0,2	0,1	0	0,9
Asco	0,1	**8,4**	1,2	0,6	1,0	0,7
Miedo	0,9	0,3	**9,9**	0,2	0,4	0,3
Alegría	0,5	0,4	0,5	**9,7**	0,4	0,5
Tristeza	0	0,5	0,9	0,9	**9,4**	0,3
Sorpresa	1,2	0,4	0,6	0,3	0,1	**9,4**

Capítulo 6

Conclusiones y trabajos futuros

En este trabajo se diseñó un sistema de reconocimiento automático de emociones en contenido multimodal. Este sistema puede resumirse en: extracción de características mediante la aplicación de PCA sobre área de ojos y boca en video; extracción de características prosódicas, espectrales y cepstrales en audio; y clasificación mediante un clasificador profundo pre-entrenado mediante la técnica de autocodificadores apilados. El sistema fue probado con la base de datos *RML Emotion Database* que cuenta con 8 individuos de distintas culturas expresando emociones en 6 idiomas diferentes.

En los experimentos se comparó el desempeño del método propuesto respecto a un clasificador tipo MLP estándar, ante distintas variantes de vectores de características. Se obtuvieron mejores resultados con el método propuesto en todas las variantes de vectores de características, lo que valida la elección de clasificación profunda frente a un MLP estándar de 3 capas. Se obtuvo un rendimiento de clasificación del 79,724 % con la arquitectura 116+100+50+6, con lo que se mejoró el rendimiento del clasificador estándar por un 3 % aproximadamente para el mismo vector de características.

Se comparó el desempeño de la clasificación multimodal contra la clasificación monomodo (audio y video). Utilizando el vector de caracterísiticas de audio con 46 coeficientes se obtuvo un desempeño del 77 % y con el vector de características de video de 85 coeficientes se obtuvo una tasa clasificación de del 75 %, mientras que con estos vectores combinados se obtuvo una tasa de clasificación superior al 78 %. Se concluye que el enfoque multimodal mejora el rendimiento de clasificación frente a la clasificación monomodal.

Se realizaron pruebas para evaluar el método propuesto frente a un único idioma expresado por individuos de diferentes culturas y en el mejor caso se obtuvo un 78 % de clasificación, resultado que es similar al obtenido frente a la clasificación multiidioma. Este resultado permite concluir que el método propuesto es robusto ante la clasificación multiidioma y multicultura.

Como trabajo futuro se podría incorporar un agrupamiento automático de emociones en base a sus características espectrales, de manera de realizar una clasificación jerárquica que permita separar las emociones más difíciles de distinguir como se realizó en [4]. Por

otro lado, sería interesante adaptar el método para que pueda ser utilizado en bases de datos de emociones *espontáneas*, para lo cual sería necesario entonces estudiar técnicas que permitan explotar la dinámica temporal en las características acústicas y de video. Además, sería de interés estudiar otras formas de fusión de la información multimodal, para comparar resultados y encontrar la que maximice la correlación de los datos.

Lista de acrónimos

A

AdaSVM *AdaBoost SVM*. 15

ANN Redes Neuronales Artificiales, del inglés *Artificial Neural Networks*. 16

ANNA Redes Neuronales Artificiales con bucle de retroalimentación, del inglés *Artificial Neural Network with a feedback loop*. 16

C

CC Cepstrum Complejo. 21

CR Cepstrum Real. 21

D

DPM Modelo de Partes Deformables, del inglés *Deformable Part Models*. 44

F

FAPs Parámetros de Animación Facial, del inglés *Facial Animation Parameters*. 14, 16

FL Puntos Faciales de Interés, del inglés *Facial Landmarks*. 44, 45

FLDA Análisis de Discriminante Lineal de Fisher, del inglés *Fisher's Linear Discriminant Analysis*. 16, 17

H

HCI Interacción Humano-Computadora, del inglés *Human–Computer Interaction*. 14

HMM Modelos Ocultos de Markov, del inglés *Hidden Markov Models*. 15, 16

HNR Relación Armónicas-Ruido, del inglés *Harmonics-to-Noise Ratio*. 15, 16

K

KCCA Análisis de Correlación de Núcleo Canónico, del inglés *Kernel Canonical Correlation Analysis*. 14

L

LDA Análisis de Discriminante Lineal, del inglés *Linear Discriminant Analysis*. 16, 17

LG Grafos Etiquetados, del inglés *Labelled Graph*. 14

LLP Patrones Lineales Largos, del inglés *Long Linear Patterns*. 16

LTI Lineal e Invariante en el Tiempo. 20, 21

M

MFCC Coeficientes Cepstrales en Escala de Mel, del inglés *Mel Frequency Cepstral Coefficients*. 15, 16, 22, 39, 40, 43, 55, 56, 58, 63, 65–67

MFHMM Modelos Ocultos de Markov Multiflujo Fusionados, del inglés Multi-stream Fused Hidden Markov Models. 16, 17

MHMM Modelos Ocultos de Markov de Multiresolución, del inglés *Multiresolution Hidden Markov Models*. 17

MLP Perceptrón Multicapa, del inglés *Multilayer Perceptron*. 15, 16, 34, 46, 59–61, 63, 64, 66, 67, 69

MLS Media del Espectro Logarítmico, del inglés *Mean Log-Spectrum*. 40, 43, 55

MS-HMM Modelos Ocultos de Markov Multietapa, del inglés *Multi Stage Hidden Markov Models*. 15

MSE Error Cuadrático Medio, del inglés *Mean Squared Error*. 47, 60–66

P

PCA Análisis de Componentes Principales, del inglés *Principal Component Analysis*. 14, 16, 22–24, 27, 28, 40, 43–46, 57, 58, 65, 69

PEFAC Algoritmo de Estimación del Pitch con Compresión de Amplitud, del inglés *Pitch Estimation Algorithm with Amplitude Compression*. 41, 42, 56

S

SNA Sistema Nervioso Autónomo. 10, 11

SNNS *Stuttgart Neural Networks Simulator*. 52, 58, 59

SNoW Red Rala de Winnows, del inglés *Sparse Network of Winnows*. 17

SNR Relación Señal Ruido. 19, 55

SVM Máquinas de Soporte Vectorial, del inglés *Support Vector Machines*. 15, 16

T

TDF Transformada Discreta de Fourier. 18, 20, 21

Bibliografía

[1] Abad, C. M. y Fernández, A. J. La expresión de la emoción a través de la conducta vocal. *Revista de psicología general y aplicada: Revista de la Federación Española de Asociaciones de Psicología* 43, 3 (1990), 289-299.

[2] Afzal, S. Affect inference in learning environments: a functional view of facial affect analysis using naturalistic data. *University of Cambridge, Computer Laboratory, Technical Report* UCAM-CL-TR-793 (2010).

[3] Albornoz, E. M., Crolla, M. B. y Milone, D. H. Recognition of emotions in speech. En: *XXXIV Conferencia Latinoamericana de Informática*. Santa Fe, Argentina, sep. de 2008, 1120-1129.

[4] Albornoz, E. M., Milone, D. H. y Rufiner, H. L. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language* 25, 3 (2011), 556-570.

[5] Aleksic, P. S. y Katsaggelos, A. K. Automatic facial expression recognition using facial animation parameters and multistream HMMs. *Information Forensics and Security, IEEE Transactions on* 1, 1 (2006), 3-11.

[6] Anderson, K. y McOwan, P. W. A real-time automated system for the recognition of human facial expressions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 36, 1 (2006), 96-105.

[7] Ayadi, M. E., Kamel, M. S. y Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572 -587. issn: 0031-3203.

[8] Banda, N. y Robinson, P. Noise Analysis in Audio-Visual Emotion Recognition (2011).

[9] Bartlett, M. S. y col. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. En: *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*. Vol. 5. 2003, 53-53.

[10] Baykal, N. Speech Emotion Recognition Using Auditory Models. Tesis doct. Middle East Technical University, 2013.

[11] Bettadapura, V. Face Expression Recognition and Analysis: The State of the Art. *CoRR* abs/1203.6722 (2012).

[12] Bourlard, H. y Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. English. *Biological Cybernetics* 59, 4-5 (1988), 291-294. issn: 0340-1200.

[13] Busso, C., Bulut, M. y Narayanan, S. *Toward effective automatic recognition systems of emotion in speech*. Oxford University Press, 2012.

[14] Busso, C. y col. Analysis of emotion recognition using facial expressions, speech and multimodal information. En: *Proceedings of the 6th international conference on Multimodal interfaces*. ACM. 2004, 205-211.

[15] Carlson, A. y col. *The SNoW learning architecture*. Inf. téc. Technical report UIUCDCS, 1999.

[16] Chang, D. F. K.-h. y Canny, J. Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense* 2011 (2011).

[17] Chang, K.-h. y col. How's my mood and stress?: an efficient speech analysis library for unobtrusive monitoring on mobile phones. En: *Proceedings of the 6th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics y Telecommunications Engineering). 2011, 71-77.

[18] Chen, T. y Rao, R. Audio-visual integration in multimodal communication. *Proceedings of the IEEE* 86, 5 (1998), 837-852. issn: 0018-9219.

[19] Chóliz, M. Psicologıa de la emoción: el proceso emocional. *Universidad de Valencia, España* (2005).

[20] Churches, O. y col. Emoticons in mind: An event-related potential study. *Social Neuroscience* 9, 2 (2014). PMID: 24387045, 196-202.

[21] Cibau, N. E., Albornoz, E. M. y Rufiner, H. L. Speech emotion recognition using a deep autoencoder. *Anales de la XV Reunión de Procesamiento de la Información y Control* (2013), 934-939.

[22] Cowie, R., McKeown, G. y Gibney, C. The challenges of dealing with distributed signs of emotion: theory and empirical evidence. En: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE. 2009, 1-6.

[23] Cowie, R. y col. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE* 18, 1 (2001), 32-80.

[24] Cowie, R. y col. *Induction techniques developed to illuminate relationships between signs of emotion and their context, physical and social*. Oxford: OUP, 2010.

[25] Deller, J. R., Hansen, J. H. L. y Proakis, J. G. *Discrete Time Processing of Speech Signals*. 1993.

[26] Devillers, L. y col. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. En: *5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*. 2006.

[27] Douglas-Cowie, E., Cowie, R. y Schroeder, M. The description of naturally occurring emotional speech. En: *Proceedings of 15th International Congress of Phonetic Sciences, Barcelona*. 2003.

[28] Eisenberg, N. y col. Relation of sympathy and personal distress to prosocial behavior: a multimethod study. *Journal of personality and social psychology* 57, 1 (1989), 55.

[29] Ekman, P. Pan-Cultural Elements in Facial Displays of Emotion. *Science* 167 (1969), 86-88.

[30] Ekman, P. Cross-cultural studies of facial expression. *Darwin and facial expression: A century of research in review* (1973), 169-222.

[31] Ekman, P. *Darwin and facial expression: A century of research in review*. Ishk, 2006.

[32] Ekman, P. Facial expression and emotion. *American Psychologist* 48, 4 (1993), 384.

[33] Ekman, P. Universals and cultural differences in facial expressions of emotion. En: *Nebraska symposium on motivation*. University of Nebraska Press. 1971.

[34] Ekman, P. y Davidson, R. J. Voluntary smiling changes regional brain activity. *Psychological Science* 4, 5 (1993), 342-345.

[35] Esteves, F., Dimberg, U. y Öhman, A. Automatically elicited fear: Conditioned skin conductance responses to masked facial expressions. *Cognition & Emotion* 8, 5 (1994), 393-413.

[36] Eyben, F., Wollmer, M. y Schuller, B. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. En: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE. 2009, 1-6.

[37] Eyben, F. y col. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces* 3, 1-2 (2010), 7-19.

[38] Fasel, B. y Luettin, J. Automatic facial expression analysis: a survey. *Pattern Recognition* 36, 1 (2003), 259 -275. issn: 0031-3203.

[39] Felzenszwalb, P. F. y col. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32, 9 (2010), 1627-1645.

[40] Ferrand, C. T. Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice* 16, 4 (2002), 480-487.

[41] Fragopanagos, N. y Taylor, J. G. Emotion recognition in human–computer interaction. *Neural Networks* 18, 4 (2005), 389-405.

[42] Frick, R. W. Communicating emotion: The role of prosodic features. *Psychological Bulletin* 97, 3 (1985), 412.

[43] Gao, L., Qi, L. y Guan, L. Selecting discriminative features with discriminative multiple canonical correlation analysis for multi-feature information fusion. En: *Biometrics Special Interest Group (BIOSIG), 2013 International Conference of the*. IEEE. 2013, 1-8.

[44] Go, H.-J. y col. Emotion recognition from the facial image and speech signal. En: *SICE 2003 Annual Conference*. Vol. 3. IEEE. 2003, 2890-2895.

[45] Gonzalez, S. y Brookes, M. PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 22, 2 (feb. de 2014), 518-530. issn: 2329-9290.

[46] Haq, S., Jackson, P. J. y Edge, J. Audio-visual feature selection and reduction for emotion classification. En: *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*. 2008.

[47] Haykin, S. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[48] Heritage, J. Distinguishing Between Non-prototypical Emotional Expressions Using a Hidden Markov Model Approach (2013).

[49] Hinton, G. E. y Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504-507.

[50] Hoch, S. y col. Bimodal fusion of emotional data in an automotive environment. En: *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. Vol. 2. IEEE. 2005, ii-1085.

[51] Huang, Y. y col. Human emotion recognition using the adaptive sub-layer-compensation based facial edge detection. En: *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE. 2013, 2876-2879.

[52] Hui, L., Dai, B.-q. y Wei, L. A pitch detection algorithm based on AMDF and ACF. En: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Vol. 1. IEEE. 2006, I-I.

[53] Izard, C. E. *The psychology of emotions*. Springer, 1991.

[54] Izenman, A. J. *Linear Discriminant Analysis*. Springer, 2008.

[55] Johnstone, T. y Scherer, K. R. Vocal communication of emotion. *Handbook of emotions* 2 (2000), 220-235.

[56] Keltner, D. y Bonanno, G. A. A study of laughter and dissociation: distinct correlates of laughter and smiling during bereavement. *Journal of personality and social psychology* 73, 4 (1997), 687.

[57] Keltner, D., Young, R. C. y Buswell, B. N. Appeasement in human emotion, social practice, and personality. *Aggressive Behavior* 23, 5 (1997), 359-374.

[58] Kessens, J. y col. Perception of synthetic emotion expressions in speech: Categorical and dimensional annotations. En: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE. 2009, 1-5.

[59] Kleinginna Jr, P. R. y Kleinginna, A. M. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion* 5, 4 (1981), 345-379.

[60] Kobayashi, V. A Hybrid Distance-based Method and Support Vector Machines for Emotional Speech Detection (2013).

[61] Kotsia, I., Buciu, I. y Pitas, I. An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing* 26, 7 (2008), 1052-1067.

[62] Levenson, R. W., Ekman, P. y Friesen, W. V. Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology* 27, 4 (1990), 363-384.

[63] Lugger, M. y Yang, B. Psychological motivated multi-stage emotion classification exploiting voice quality features. *Speech Recognition, In-Tech* (2008).

[64] Martínez, C. E. y col. An approach to robust phoneme classification with auditory cortical representation for speech. En: *XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009)*. Rosario, Argentina, sep. de 2009, 411-416.

[65] Meghjani, M., Ferrie, F. y Dudek, G. Bimodal information analysis for emotion recognition. En: *Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE. 2009, 1-6.

[66] Mehrabian, A. Communication without words. *Psychology Today* 2, 4 (1968), 53-56.

[67] Michel, P. y El Kaliouby, R. Real time facial expression recognition in video using support vector machines. En: *Proceedings of the 5th international conference on Multimodal interfaces*. ACM. 2003, 258-264.

[68] MN, H., Hariharan, M. y Sazali, Y. Speech emotion recognition using kNN classifier (2012).

[69] Morrison, D., Wang, R. y De Silva, L. C. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication* 49, 2 (2007), 98-112.

[70] Nöth, E. y col. On the use of prosody in automatic dialogue understanding. *Speech Communication* 36, 1 (2002), 45-62.

[71] Nunes, A. M. B. Cross-linguistic and Cultural effects on the perception of emotions (2012).

[72] O'Connor, J. D. y Gordon, F. A. *Intonation of colloquial English, second ed.* Longman, 1973.

[73] Oflazoglu, C. y Yildirim, S. Recognizing emotion from Turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing* 2013, 1 (2013), 26.

[74] Oppenheim, A. V., Willsky, A. S. y Young, I. T. *Signals and systems*. Prentice-Hall, 1983.

[75] Pantic, M. y Patras, I. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. En: *Systems, Man and Cybernetics, 2005 IEEE International Conference on*. Vol. 4. IEEE. 2005, 3358-3363.

[76] Pantic, M. y Rothkrantz, L. J. M. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22, 12 (2000), 1424-1445.

[77] Pardàs, M. y Bonafonte, A. Facial animation parameters extraction and expression recognition using Hidden Markov Models. *Signal Processing: Image Communication* 17, 9 (2002). Image Processing for 3-D Imaging, 675 -688. issn: 0923-5965.

[78] Regenbogen, C. y col. Multimodal human communication — Targeting facial expressions, speech content and prosody. *NeuroImage* 60, 4 (2012), 2346 -2356. issn: 1053-8119.

[79] Regenbogen, C. y col. The differential contribution of facial expressions, prosody, and speech content to empathy. *Cognition and Emotion* 26, 6 (2012). PMID: 22214265, 995-1014.

[80] Rosenberg, E. L. y col. Linkages between facial expressions of anger and transient myocardial ischemia in men with coronary artery disease. *Emotion* 1, 2 (2001), 107.

[81] Rufiner, H. L. *Análisis y modelado digital de la voz. Técnicas recientes y aplicaciones*. Universidad Nacional del Litoral, 2009.

[82] Scherer, K. R. Expression of emotion in voice and music. *Journal of Voice* 9, 3 (1995), 235 -248. issn: 0892-1997.

[83] Scherer, K. R. Nonlinguistic vocal indicators of emotion and psychopathology. En: *Emotions in personality and psychopathology*. Springer, 1979, 493-529.

[84] Scherer, K. R., Johnstone, T. y Klasmeyer, G. Vocal expression of emotion. *Handbook of affective sciences* (2003), 433-456.

[85] Schroder, M. Expressing degree of activation in synthetic speech. *Audio, Speech, and Language Processing, IEEE Transactions on* 14, 4 (2006), 1128-1136.

[86] Schröder, M. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. En: *Affective dialogue systems*. Springer, 2004, 209-220.

[87] Schubiger, M. *English intonation*. Niemeyer, 1958.

[88] Schuller, B. y col. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53, 9 (2011), 1062-1087.

[89] Schuller, B. y col. Speaker independent speech emotion recognition by ensemble classification. En: *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE. 2005, 864-867.

[90] Shama, K., Cholayya, N. U. y col. Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *EURASIP Journal on Applied Signal Processing* 2007, 1 (2007), 50-50.

[91] Sneddon, I. y col. Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour. *PloS one* 6, 2 (2011), e14679.

[92] Snel, J. y col. A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus (2012).

[93] Song, T. y col. Recognizing human emotional state via SRC in Fractional Fourier Domain. En: *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*. Vol. 3. IEEE. 2012, 1583-1586.

[94] Stein, B. E. y Meredith, M. A. *The merging of the senses.* The MIT Press, 1993.

[95] Tie, Y. y Guan, L. Human emotion recognition using a deformable 3D facial expression model. En: *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*. 2012, 1115-1118.

[96] Trentin, E. y Gori, M. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing* 37, 1–4 (2001), 91 -126. issn: 0925-2312.

[97] Truong, K. y Van Leeuwen, D. An 'open-set'detection evaluation methodology for automatic emotion recognition in speech. En: *Workshop on Paralinguistic Speech-between models and data*. 2007, 5-10.

[98] Turk, M. y Pentland, A. Eigenfaces for recognition. *Journal of cognitive neuroscience* 3, 1 (1991), 71-86.

[99] Uřičář, M., Franc, V. y Hlaváč, V. Detector of Facial Landmarks Learned by the Structured Output SVM. En: *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*. (Rome, Italy). Ed. por Csurka, G. y Braz, J. Vol. 1. SciTePress — Science y Technology Publications, Portugal, 24-26 de feb. de 2012, 547-556. isbn: 978-989-8565-03-7.

[100] Vinciarelli, A. y Pantic, M. Techware: www. sspnet. eu: A Web Portal for Social Signal Processing [Best of the Web]. *Signal Processing Magazine, IEEE* 27, 4 (2010), 142-144.

[101] Wang, Y. y Guan, L. Recognizing human emotion from audiovisual information. En: *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. Vol. 2. IEEE. 2005, ii-1125.

[102] Wang, Y., Guan, L. y Venetsanopoulos, A. N. Kernel Cross-Modal Factor Analysis for Information Fusion With Application to Bimodal Emotion Recognition. *Multimedia, IEEE Transactions on* 14, 3 (2012), 597-607.

[103] Wang, Y., Guan, L. y Venetsanopoulos, A. N. Kernel cross-modal factor analysis for multimodal information fusion. En: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE. 2011, 2384-2387.

[104] Wang, Y. y col. Kernel fusion of audio and visual information for emotion recognition. En: *Image Analysis and Recognition*. Springer, 2011, 140-150.

[105] Wiley, E. O. y Lieberman, B. S. *Phylogenetics: Theory and practice of phylogenetic systematics*. John Wiley & Sons, 2011.

[106] Wu, C.-H., Yeh, J.-F. y Chuang, J. Emotion perception and recognition from speech. En: *Affective Information Processing*. Springer, 2009, 93-110.

[107] Xie, Z. y Guan, L. Multimodal Information Fusion of Audio Emotion Recognition Based on Kernel Entropy Component Analysis. En: *Multimedia (ISM), 2012 IEEE International Symposium on*. IEEE. 2012, 1-8.

[108] Xie, Z. y Guan, L. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. En: *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE. 2013, 1-6.

[109] Yuasa, M., Saito, K. y Mukawa, N. Emoticons Convey Emotions Without Cognition of Faces: An fMRI Study. En: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '06. ACM, Montreal, Quebec, Canada, 2006, 1565-1570. isbn: 1-59593-298-4.

[110] Yumoto, E., Gould, W. J. y Baer, T. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The journal of the Acoustical Society of America* 71, 6 (1982), 1544-1550.

[111] Zeng, Z. y col. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31, 1 (2009), 39-58.

[112] Zeng, Z. y col. Audio-visual affect recognition. *Multimedia, IEEE Transactions on* 9, 2 (2007), 424-428.

[113] Zeng, Z. y col. Audio-visual affect recognition in activation-evaluation space. En: *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE. 2005, 4-pp.

[114] Zeng, Z. y col. Audio-visual affect recognition through multi-stream fused HMM for HCI. En: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 2. IEEE. 2005, 967-972.

[115] Zeng, Z. y col. Audio-visual spontaneous emotion recognition. En: *Artifical Intelligence for Human Computing*. Springer, 2007, 72-90.

[116] Zeng, Z. y col. Bimodal HCI-related affect recognition. En: *Proceedings of the 6th international conference on Multimodal interfaces*. ACM. 2004, 137-143.

[117] Zeng, Z. y col. Training combination strategy of multi-stream fused hidden Markov model for audio-visual affect recognition. En: *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM. 2006, 65-68.

[118] Zhang, L. y col. Recognizing smile emotion based on Fractional Fourier Transform. En: *Image and Signal Processing (CISP), 2011 4th International Congress on*. Vol. 2. IEEE. 2011, 940-944.

[119] Zheng, W. y col. Facial expression recognition using kernel canonical correlation analysis (KCCA). *Neural Networks, IEEE Transactions on* 17, 1 (2006), 233-238.