Universidad Autónoma Metropolitana – Iztapalapa
División de Ciencias Básicas e Ingeniería

RECONOCIMIENTO DE EMOCIONES UTILIZANDO TÉCNICAS DE APRENDIZAJE MAQUINAL

Idónea Comunicación de Resultados
que presenta
Máximo Eduardo Sánchez Gutiérrez
Para obtener el grado de
Maestro en Ciencias y Tecnologías de la Información

Asesores: M. I. Fabiola M. Martínez Licona
Dr. John Goddard Close

Jurado Calificador:

Presidente:	DR. HUGO JAIR ESCALANTE BALDERAS	INAOE
Secretario:	M.I. FABIOLA M. MARTÍNEZ LICONA	UAM-I
Vocal:	DR. PEDRO PABLO GONZÁLEZ PÉREZ	UAM-C

México D.F. octubre 2013

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

NOMENCLATURA

ANN	Redes Neuronales Artificiales
BM	Máquina de Boltzmann
CART	Árboles de Clasificación y Regresión
CD	Divergencia Contrastiva
DBN	Redes de Creencia Profunda
DL	Aprendizaje Profundo
DT	Árboles de Decisión
GMM	Modelos de Mezcla de Gaussianas
HMM	Modelos Ocultos de Markov
ISCA	International Speech Communication Association
KNN	K-Vecinos más Cercanos
LDA	Análisis Discriminante Lineal

LFPC	Coeficientes de Potencia de la Frecuencia Logarítmica
LLD	Descriptores de Bajo Nivel
LPC	Codificación Lineal Predictiva
LPCC	Predicción Linear de Coeficientes Cepstrales
MFCC	Coeficientes Cepstrales en las frecuencias de Mel
PCA	Análisis de Componentes Principales
QR code	Códigos de Respuesta Rápida
RBM	Máquina Restringida de Boltzmann
SERS	Sistema de Reconocimiento de Emociones en el Habla
SVM	Máquinas de Soporte Vectorial
UPC	Universidad Politécnica de Catalunya
ZCR	Tasa de Cruces por Cero

RESUMEN

La señal de habla se caracteriza por su alta variabilidad, pues su producción queda condicionada por la ubicación y movimiento de los elementos en la cavidad oral y el rostro, y por variantes en parámetros como el acento regional, la condición social o el estilo personal. La expresión de emociones es otro de los elementos que enriquecen la comunicación humana, incluso se ha establecido que las palabras por sí mismas no aportan el significado completo del mensaje para un escucha, por lo que el análisis de los componentes paralingüísticos como la prosodia, la calidad de la voz, el ritmo e incluso las emociones con las que son dichas las palabras se ha vuelto importante.

Con esto en mente, es que este trabajo aborda el problema del análisis y reconocimiento de emociones a partir de la señal del habla, utilizando técnicas de extracción de datos y aprendizaje maquinal. Con este fin se utilizaron métodos de computación afectiva, los cuales, desde la perspectiva de la interacción hombre-máquina, se enfocan en la detección, análisis y reconocimiento automático de las emociones a partir del comportamiento humano. Otra área importante es la biología, ésta nos dice que el cerebro funciona de forma profunda, dicha naturaleza, a la que llamaremos jerárquica, proviene de la

observación de que generalmente las capas superiores representan conceptos cada vez más abstractos. Este paradigma profundo, plantea la hipótesis de que, con el fin de aprender las representaciones de alto nivel de los datos, se necesita una jerarquía de representaciones intermedias. Al mismo tiempo se ha sugerido que las arquitecturas profundas, son mucho más eficientes en términos de los elementos computacionales requeridos, que las arquitecturas simples o de poca profundidad, incluyendo los Modelos Ocultos de Markov, las Redes Neuronales con una sola capa oculta y las Máquinas de Soporte Vectorial entre otras.

Es por esto, que el objetivo principal fue desarrollar un sistema de reconocimiento de emociones a partir de la señal del habla, mediante el uso de Redes de Creencia Profunda y Máquinas Restringidas de Boltzmann que clasificará señales de audio caracterizadas por un grupo de elementos propuestos. Para evaluar este sistema se desarrollaron distintos experimentos, que permitieron realizar una comparación con otros clasificadores ampliamente utilizados. Los resultados mostraron que nuestra propuesta se desempeña comparativamente mejor.

De los resultados también se desprende una primera aproximación a la interpretación del funcionamiento de los distintos niveles de abstracción en la arquitectura profunda que hemos utilizado. La idea general detrás de la interpretación, es que si conocemos los mecanismos que intervienen en la clasificación de emociones mediante sistemas profundos, podremos determinar los parámetros y características que resulten convenientes para resolver el problema y con ello aumentar el desempeño del clasificador.

CAPÍTULO 1

INTRODUCCIÓN

En este capítulo se da una panorámica general, así como la motivación que existe en distintos ámbitos para la creación de un sistema automático de reconocimiento de emociones a partir de la voz, basado en la teoría de las Máquinas Restringidas de Boltzmann (RBM), y Redes de Creencia Profunda (DBN) como sistema de clasificación [2, 3]. También se describen los objetivos y la estructura del documento.

Los Sistemas de Reconocimiento de Emociones en el Habla (SERS) requieren principalmente de un conjunto de grabaciones de audio sobre las cuales trabajar, una selección de características que describan las emociones y, por supuesto, un método de clasificación. Cuando se logran satisfacer estos requerimientos, se cuenta entonces con un sistema que, basado en una muestra de voz y a través de los procedimientos que se irán desarrollando a lo largo de este trabajo, es capaz de determinar si ésta corresponde a alguna emoción. En las siguientes secciones se exponen los planteamientos al respecto.

1.1. Contexto e identificación de la problemática

El objetivo principal, al emplear el reconocimiento de emociones en la voz, es el de hacer que las respuestas de los sistemas computacionales se adapten a las emociones del hablante. Sin embargo, reconocer toda la gama de emociones que puede presentar el ser humano no es tarea sencilla. Se han logrado identificar más de 300 emociones humanas [4, 5], algunas de ellas se exhiben en la Tabla 1.1:

Aburrido	Ambivalente	Atónito	Desobediente	Impresionado
Aceptar	Animado	Aventurero	Disgustado	Insatisfecho
Afectuoso	Antagónico	Avergonzado	Disgustado	Molesto
Agradable	Apático	Calmado	Divertido	Perplejo
Agresivo	Aprehensivo	Cauteloso	Enojado	Presumido
Alegre	Asombrado	Desconfiado	Estupefacto	Previsor
Amargado	Atento	Desinteresado	Eufórico	Vacío

Tabla 1.1: *Ejemplo de algunas emociones identificadas por J.D. O'Connor y Gordon Frederick Arnold en [5], 1973*

Existen diversas cuestiones que necesitan ser abordadas con cuidado, ya que representan retos importantes para la correcta clasificación de las emociones. Tal es el caso de las 'características' del habla, donde es importante identificar la duración de las palabras, la cadencia, el volumen, la velocidad, etc.

Otro reto que se presenta es el de la variabilidad entre oraciones, estilos del habla, e incluso, hablantes; como sabemos, existen diferencias en la forma en la que se expresan las personas de distintas regiones de un mismo país. Incluso se da el caso de distintas formas de expresión en los diferentes estratos

sociales, lo que indica que los hablantes expresan emociones de acuerdo con su cultura y medio ambiente [6]; además, nuestro estado de ánimo no es constante ni periódico, pues durante un día podemos oscilar entre éstos, es decir, una oración puede ser expresada con una mezcla de diferentes emociones que tienen un tiempo de permanencia diferente, por ejemplo el enojo suele tener una duración menor a la depresión. Por estas razones se ha sugerido que las palabras por sí mismas, no aportan el significado completo del mensaje para un escucha, por lo que el análisis de los componentes paralingüísticos como la prosodia, la calidad de la voz, el ritmo e incluso las emociones con las que son dichas las palabras se ha vuelto importante [7].

Algo que merece la pena dejar claro, es que las emociones no tienen una definición teórica aceptada por todos y por ende lo que pudiera significar alegría para unos no lo será para otros. Lo que sí sabemos sobre las emociones es que se puede medir, en la activación de la voz, la energía requerida para expresar una emoción. Haciendo uso de esto podemos explicar el concepto de 'emoción en el habla' como un estado, en donde la intensidad empleada para pronunciar una frase, ha sobrepasado un cierto límite.

Se han ofrecido distintas formulaciones de esa idea; pensando en las emociones primarias o básicas como puntos cardinales, la 'rueda de emoción' [8] que se muestra en la Figura 1.1 o el sistema Feeltrace [9] que se muestra en la Figura 1.2 proponen una clasificación. Incluso se han ofrecido aproximaciones como es el caso del 'cubo de emociones', Figura 1.3, donde las emociones son clasificadas mediante la presencia de algunas substancias químicas en el cerebro [10]. Estas teorías son semejantes a una paleta de colores donde los colores primarios se mezclan para producir nuevas combinaciones. En este trabajo utilizamos las seis emociones primarias o arquetípicas: Enojo, Disgusto, Miedo, Alegría, Tristeza y Sorpresa que combinadas pueden producir otra gama de emociones además de la neutral [11, 12].

En las Figuras 1.1 y 1.2 podemos notar que el enojo y la alegría se corresponden con una alta activación; ésta es una de las características del habla que se utilizan para determinar la emoción que se presenta en la señal de la voz, estas propiedades se revisan en la Sección 2.2 donde se da una descripción de los conjuntos de características prosódicas, espectrales y de calidad del habla.

Existen distintas áreas en las que el reconocimiento de emociones en la voz podría ser aplicado, por lo que se han planteado diferentes aproximaciones al problema. Estas propuestas utilizan ya sea información obtenida mediante video o grabaciones de audio; para el caso de las que utilizan el video como fuente de información se utiliza la detección de micro expresiones en el rostro [13], mientras que para aquellas que utilizan la señal de la voz, como en nuestro caso, se aplican algunas de las técnicas para la extracción de características que se detallarán más adelante en este trabajo, incluso existen aquellas aproximaciones en las que se propone reconocer emociones en piezas musicales [14, 15]. A continuación se dan algunos ejemplos de estas áreas que motivan aún más el estudio de este tema.

1.2. Motivación

El reconocimiento de emociones en el habla es útil en áreas que se benefician de la interacción hombre-máquina, ejemplo de esto es la generación de voces que tienen un énfasis emocional adecuado. Hay una estrecha relación entre el tema de generación de voces y el del reconocimiento de emoción en el habla. Las técnicas para aprender las sutilezas del 'habla emocionada' pueden proporcionar una forma de generar voces con una carga emocional convincente. Un caso especial surge con las técnicas de compresión, donde existe la posibilidad de extraer información acerca de la emoción, intentan-

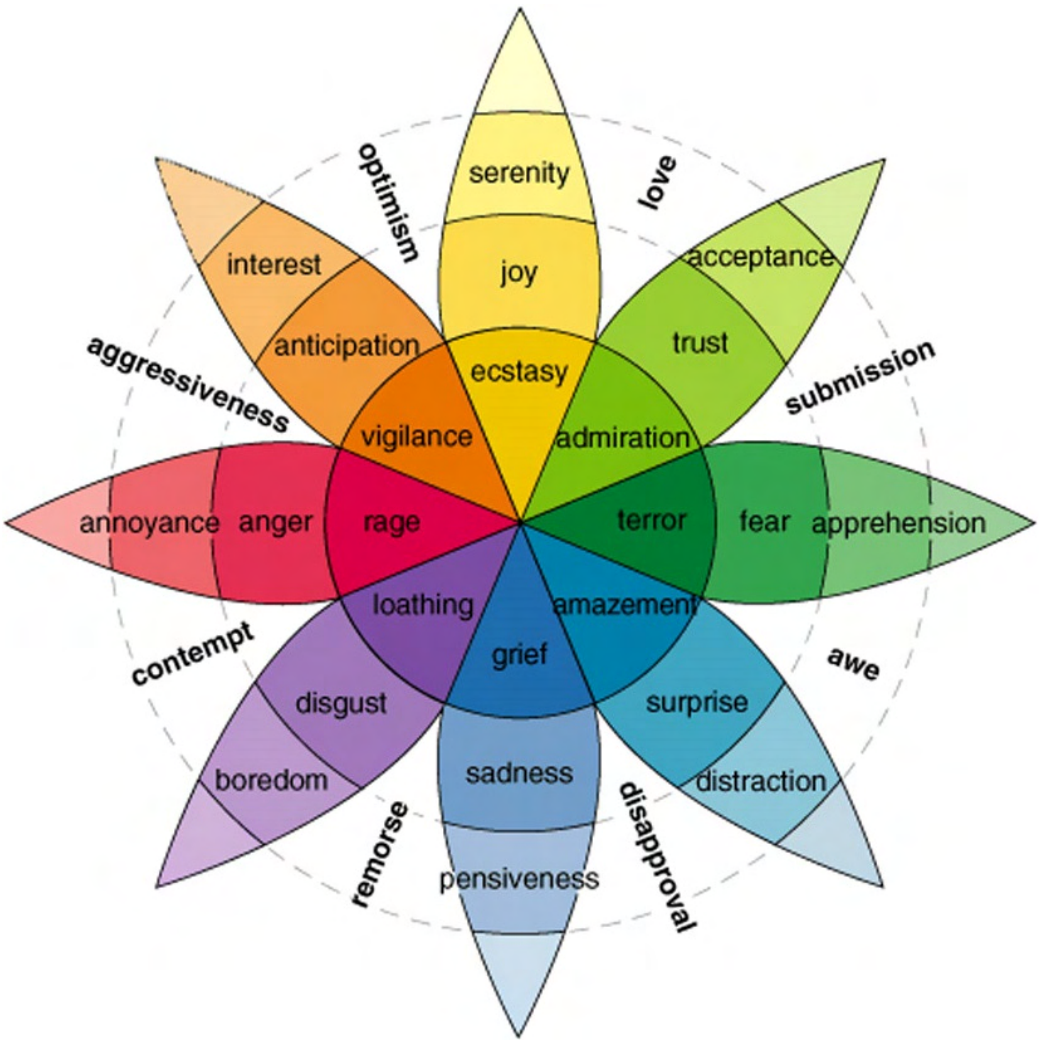

Figura 1.1: *Rueda de emociones por Robert Plutchik, tomada de [8], 2001*

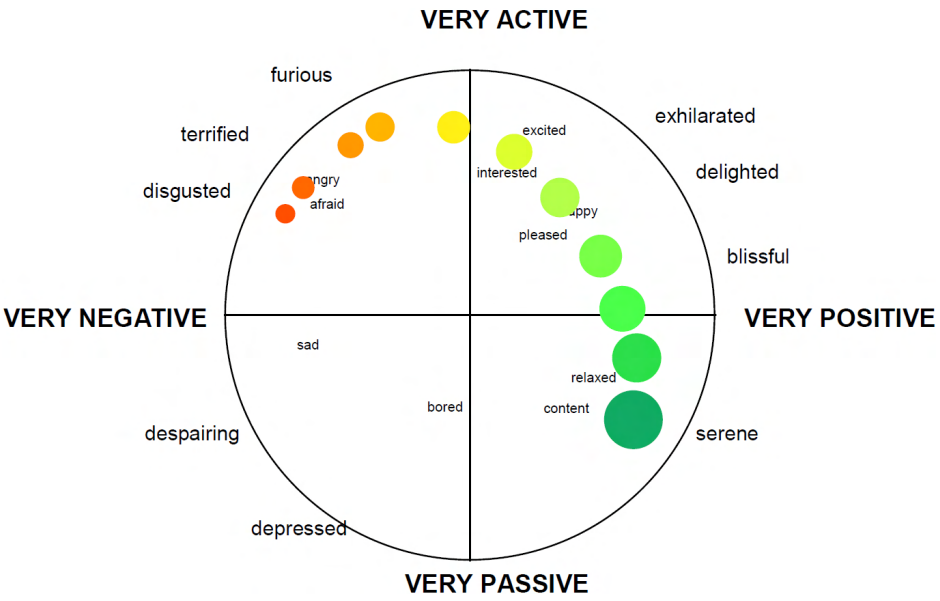

Figura 1.2: *Sistema Feeltrace tomado de Schröder Marc et al., [9] 2000*

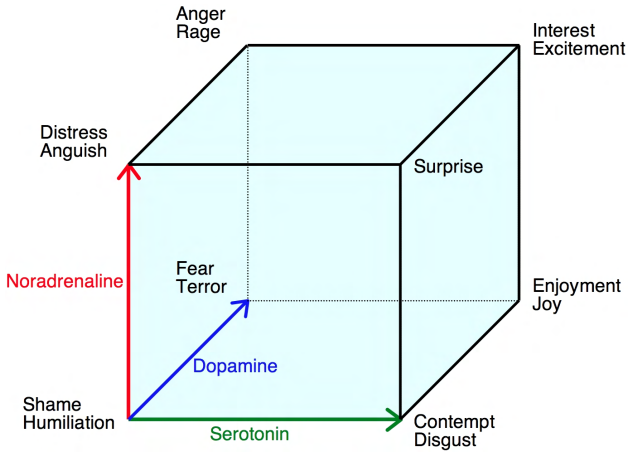

Figura 1.3: *Cubo de emociones según H Lövheim, tomado de [10], 2012*

do transmitirla para después poder resintetizarla. Si esta re-síntesis fallase al reconstruir la información acerca de la emoción, el resultado podría ser sumamente engañoso.

La enseñanza es otra área en la que resulta de mucha utilidad, no hace falta mucho para darse cuenta que un 'tutor' se beneficiaría del reconocimiento de emociones para saber si el usuario encuentra los ejemplos aburridos o interesantes [16], tampoco es difícil ver cómo de manera similar un tutor puede expandir sus capacidades hacia un asistente médico personal [17], un centro de información o incluso un recepcionista que reconozca si el usuario con el que interacciona encuentra adecuada la información que se le ofrece [18]. Continuando con la idea de un asistente médico personal, resulta de vital importancia que éste pudiera monitorear el estado anímico del paciente y alertar si se ha producido un cambio emocional hacia la depresión, incluso podríamos imaginar una situación de negociación en la que el asistente notifique sobre las emociones que pudiesen intervenir en la toma de decisiones [19].

Un área en la que el reconocimiento de emociones tendría un alto impacto es en la industria del entretenimiento, donde existe una gran posibilidad de que se aproveche este aprendizaje para desarrollar mascotas, muñecas y juegos que respondan al estado de ánimo del usuario [20].

Con todas estas áreas en las que el reconocimiento de emociones puede ser aplicado, resulta de gran importancia abordar el tema haciendo uso de técnicas novedosas y que han mostrado ser efectivas, como la del aprendizaje profundo, que ha logrado producir mejores resultados que otras no profundas [2, 3]. Con esto en mente hemos planteado los distintos objetivos que se describen a continuación.

1.3. Hipótesis

Este trabajo plantea como hipótesis que, hacer uso de las máquinas restringidas de Boltzmann en una arquitectura profunda, permite la correcta clasificación de emociones en el habla.

1.4. Objetivos

Objetivo general:

Desarrollar un sistema de reconocimiento de emociones a partir de la señal del habla mediante el uso de Redes de Creencia Profunda y Máquinas Restringidas de Boltzmann.

Objetivos particulares:

- Seleccionar los componentes acústicos, frecuenciales y paralingüísticos de la señal del habla que se puedan utilizar en un sistema de clasificación de emociones.
- Desarrollar un sistema de clasificación para el reconocimiento de estados emocionales en el habla mediante Redes de Creencia Profunda y Máquinas Restringidas de Boltzmann.
- Evaluar el desempeño de este sistema mediante la comparación con otros métodos tradicionales de clasificación.

1.5. Metodología

En esta sección se expondrán de manera general los componentes que conforman el sistema propuesto en este trabajo y, al mismo tiempo, se discutirán las decisiones tomadas. En la Figura 1.4 se muestra un esquema general de los sistemas de reconocimiento de emociones en el habla (SERS).

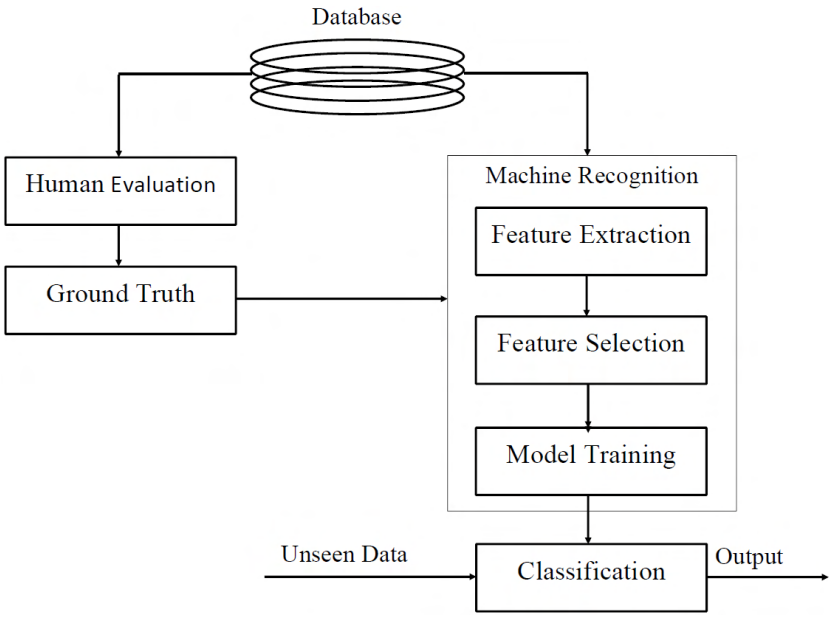

Figura 1.4: *Sistema de Reconocimiento de Emociones en el Habla tomado de Ali Hassan [21], 2012*

El primer componente y el que creemos es el requisito más importante para un algoritmo automático de clasificación es la base de datos o fuente de información.

1.5.1. Base de datos

Existen tres técnicas mediante las cuales se pueden obtener las representaciones emocionales en el habla para la realización de las bases de datos: el habla simulada en la que las emociones son deliberadamente habladas por profesionales, el habla natural o espontánea que se obtiene de grabar en un ambiente de la vida real y el habla inducida en la que los hablantes son puestos en situaciones controladas que les permiten expresarse con libertad.

La base de datos que hemos escogido es del primer tipo, de habla simulada por actores. Esta elección fue tomada pensando en las ventajas que este método de recopilación de información proporciona:

- Las representaciones que se producen son intensas y generalmente prototípicas, por lo que la clasificación posterior es mucho más fácil.
- Los estudios de grabación permiten grabar audio de alta calidad y evitan problemas en el procesamiento de las señales de voz como la reverberación o el ruido.
- Se puede garantizar una distribución equilibrada de las emociones, por lo tanto, el problema de escasez de datos se puede mitigar.
- Los datos pueden ser recopilados en un tiempo relativamente corto.
- El post-procesamiento de los datos es bastante simple en comparación con las grabaciones del habla espontánea, ya que no es necesario el etiquetado posterior de las emociones pues éste se conoce de antemano.

Es por esto que en este trabajo se ha utilizado INTERFACE [22], que es una base de datos creada en el Center for Language and Speech Technologies and Applications (TALP), de la Universidad Politécnica de Catalunya (UPC), con el propósito de estudiar el habla con emociones y la síntesis de voz. Las frases que contiene expresan seis emociones más cinco variaciones

de neutral en español. En el Capítulo 5 se puede encontrar más información acerca de esta base.

El siguiente requisito importante es el de la extracción y selección de características del habla que puedan caracterizar correctamente las emociones.

1.5.2. Conjuntos de datos

De la base de datos hemos seleccionado un conjunto reducido de tres emociones del hablante femenino con el objetivo de tener un mayor control sobre los experimentos mediante la reducción de variables, esta reducción nos permitió tener una mejor idea sobre el comportamiento del sistema así como de los parámetros utilizados en el entrenamiento. A éste subconjunto de emociones pertenecen, la neutral, la alegría y la tristeza como se discute en el Capítulo 6.

Pensando en la extensión a más emociones y distintos hablantes fue que en el Capítulo 8 se muestran los resultados de la experimentación con las emociones: enojo, disgusto, miedo, alegría, sorpresa, tristeza y neutral para dos hablantes, femenino y masculino.

1.5.3. Características del habla

Un gran número de estas propiedades se han propuesto para reconocer los estados emocionales en el habla, en los Capítulos 2 y 3 se puede encontrar información más detallada al respecto mientras que en el Anexo B se puede encontrar una lista más puntual de características.

Las características utilizadas en nuestro sistema son 30:

- 12 MFCCs y su primera derivada

- Promedio de F_0 y su primera derivada
- Promedio de los cruces por cero y su primera derivada
- Energía y su primera derivada

Estas características fueron seleccionadas experimentalmente, es decir, se comenzó a experimentar con la *frecuencia fundamental* y con la ayuda de otros trabajos sobre el reconocimiento de emociones, que se muestran en el Anexo A, se fueron incorporando incrementalmente más propiedades.

Cabe mencionar que el problema de la selección de atributos es un tema difícil de tratar, sobretodo cuando existen superconjuntos con más de mil propiedades correlacionadas como se puede ver en el Anexo B. Es por esto que otro tipo de experimentación fue llevada a cabo mediante la herramienta WEKA [23].

Con esta herramienta computacional se llevaron acabo distintos experimentos de selección de atributos que entregaron varios subconjuntos de características utilizados para realizar las pruebas descritas en el Capítulo 6. No obstante esta experimentación, los porcentajes de error se mantuvieron por encima de los obtenidos con las 30 características antes mencionadas.

1.5.4. Clasificador

Aunque se cuenta con numerosos estudios de clasificación de emociones en la voz, como se puede ver en el Capítulo 2 y en el Anexo A, existen muy pocos que combinan las arquitecturas profundas y las máquinas restringidas de Boltzmann (RBM), ambas explicadas en el Capítulo 4.

En este trabajo eso es lo que se propone, hacer uso de modelos jerárquicos de datos con el fin de aprender las representaciones intermedias y, median-

te el uso de una RBM en la capa superior que funcione como clasificador, identificar las emociones presentadas.

1.5.5. Experimentación y resultados

Los porcentajes de clasificación obtenidos en las pruebas realizadas al sistema con tres emociones se pueden ver en la Figura 6.3 mientras que los resultados con siete se pueden ver en la 8.10, estos porcentajes de error fueron de 2,51 % y 18,37 % respectivamente.

Estos resultados muestran que, con los parámetros correctos, es posible crear un sistema de reconocimiento de emociones en el habla que alcance mejores resultados que los que se presenta en la Sección 5.3.

1.6. Estructura del documento

Como se expuso en la sección anterior, nuestra propuesta consta de tres etapas generales para el reconocimiento de emociones en el habla. Estas etapas nos proporcionan la secuencia de los capítulos de este trabajo:

- Extracción de características de los audios
- Entrenamiento del sistema
- Clasificación de emociones

Capítulo 2. Se da un panorama general del funcionamiento de un SERS mediante la descripción de bases de datos, características del habla y clasificadores.

Capítulo 3. Se da un panorama de las características del habla y su extracción, también se introducen las características que son parte clave en este trabajo.

Capítulo 4. Se introduce el tema principal del trabajo, las Redes de Creencia Profunda y las Máquinas Restringidas de Boltzmann. También se explica el funcionamiento de éstas y los algoritmos de entrenamiento.

Capítulo 5. Se describe la base de datos de habla con emociones con la que se entrenó y probó nuestro trabajo.

Capítulo 6. Se muestran los resultados obtenidos de la clasificación, la elección de los parámetros que llevaron a dichos resultados y una comparación con otros clasificadores.

Capítulo 7. Se concluye el trabajo realizado en los capítulos anteriores comparándolo con los objetivos antes descritos.

Capítulo 8. Se discuten los resultados obtenidos y se comenta la primera aproximación de interpretación del aprendizaje en las capas de la DBN, también se da un primer resultado de la experimentación realizada con un conjunto de mociones más amplio así como las pruebas con otra base de datos.

CAPÍTULO 2

SISTEMAS DE RECONOCIMIENTO DE EMOCIONES EN EL HABLA

Desde hace tiempo se ha investigado la influencia de las emociones en el habla, dando como resultado algunos ejemplos de emociones prototípicas [8, 9, 10, 11, 12], al mismo tiempo las ciencias computacionales se han visto más involucradas en el tema del reconocimiento automático de las emociones por lo cual se han empezado a utilizar técnicas de identificación de patrones para clasificarlas, dando como resultado la aparición de los Sistemas de Reconocimiento de Emociones en el Habla (SERS). Estos son un caso particular de un sistema que puede tomar como entrada una señal de voz y, mediante el uso de la información extraída de la muestra, determinar el estado emocional del hablante.

En este capítulo se expondrán los componentes que conforman un sistema de este tipo, en la Figura 2.1 se muestra un esquema general de estos sistemas que en las siguientes secciones iremos describiendo. El primero de ellos y el que creemos es el requisito más importante para un algoritmo automático de clasificación es la base de datos o fuente de conocimiento.

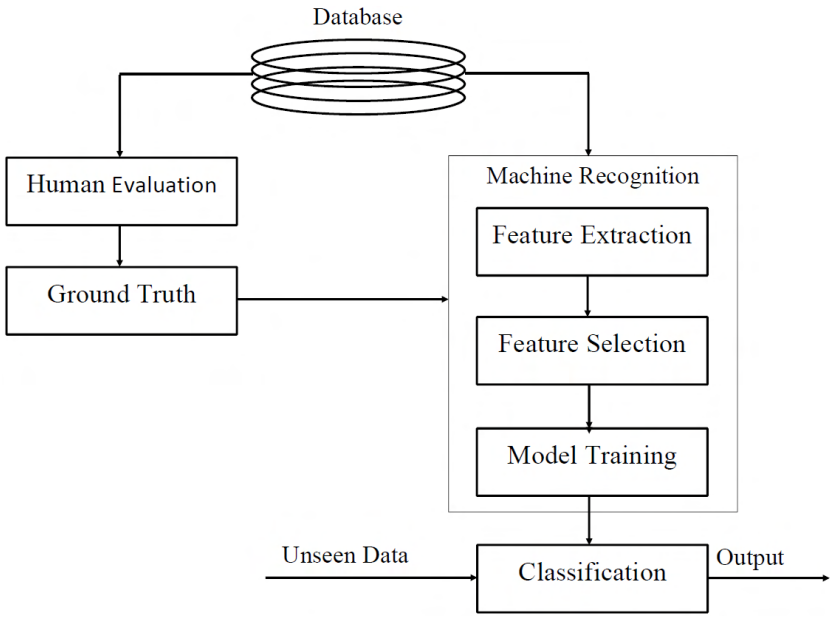

Figura 2.1: *Sistema de Reconocimiento de Emociones en el Habla tomado de Ali Hassan [21], 2012*

2.1. Bases de datos

Existen tres técnicas mediante las cuales se pueden obtener las representaciones emocionales en el habla para la realización de las bases de datos: el habla simulada en la que las emociones son deliberadamente habladas por profesionales, el habla natural o espontánea que se obtiene de grabar en un ambiente de la vida real y el habla inducida en la que los hablantes son puestos en situaciones controladas que les permiten expresarse con libertad.

Muchas de las bases de datos disponibles en la actualidad consisten en elementos emocionales recreados por actores, el uso de estos expertos se basa en las ventajas intrínsecas de este método de recopilación de información:

- Las representaciones que se producen son intensas y generalmente prototípicas, por lo que la clasificación posterior es mucho más fácil.
- Los estudios de grabación permiten grabar audio de alta calidad y evitan problemas en el procesamiento de las señales de voz como la reverberación o el ruido.
- Se puede garantizar una distribución equilibrada de las emociones, por lo tanto, el problema de escasez de datos se puede mitigar.
- Los datos pueden ser recopilados en un tiempo relativamente corto.
- El post-procesamiento de los datos es bastante simple en comparación con las grabaciones del habla espontánea, ya que no es necesario el etiquetado posterior de las emociones pues éste se conoce de antemano.

Las investigaciones en esta área se han centrado en los diferentes componentes que conforman el corpus de la base de datos [12, 24, 25, 26], algunas han contribuido mediante la creación de bases de datos que contienen oraciones cargadas emocionalmente, mientras que otras han trabajado en las características que se pueden extraer de las oraciones y que dan mejor información sobre el estado emocional del hablante. En el Anexo A se puede encontrar una revisión de algunas de las bases de datos más conocidas y los resultados obtenidos tras haberlas utilizado para el reconocimiento de emociones.

El siguiente requisito importante es el de la extracción y selección de características del habla que puedan caracterizar correctamente las emociones.

2.2. Características del habla

Un gran número de estas propiedades se han propuesto para reconocer los estados emocionales en el habla, en los siguientes párrafos veremos que estas

características se pueden clasificar en prosódicas, espectrales y de calidad en la voz; además, en el Capítulo 3 se revisan las técnicas tradicionales para la extracción de las propiedades del habla que son utilizadas por nuestro sistema.

Una lista más puntual de características se puede ver en el Anexo B en el que se detallan las propuestas en los años 2009 y 2010 de los concursos que versaron sobre cuáles serían las que mejor describirían las emociones. Estos concursos se llevaron a cabo por la International Speech Communication Association (ISCA).

2.2.1. Características prosódicas

La prosodia estudia la producción de las palabras desde el punto de vista fonético-acústico hasta la variación de la frecuencia fundamental, la duración y la intensidad; se puede decir que se divide en dos aspectos: el aspecto suprasegmental, es decir, el que trata la entonación de la frase en su conjunto y, el aspecto que controla la melodía y los fenómenos locales de coarticulación y acentuación [27].

Las propiedades prosódicas son atribuidas a segmentos de habla cuya duración es mayor a la de fonemas, como por ejemplo sílabas, palabras o frases. A este conjunto de características pertenecen el tono, el volumen, la velocidad, la duración, las pausas entre palabras y el ritmo, pues son las que permiten discernir entre el 'Sí, claro' de afirmación y el 'Sí, claro' de ironía.

En general, y dado que la percepción de estos rasgos es variable entre individuos, no tienen un equivalente único en la señal de la voz, aunque existen algunos que están altamente relacionados con características que pueden ser extraídas de la señal, tal es el caso de la Frecuencia Fundamental (F_0) que lo está con el tono.

Las características de F_0 están dadas por el cambio de sus valores a través del tiempo dentro de una palabra o frase, por lo que si se desea conocer las propiedades que definen todo el segmento hablado, se pueden aplicar ciertas funciones estadísticas al conjunto de datos. Algunas de las más comunes son la media, la mediana, el máximo, el mínimo, la desviación estándar, y el rango o diferencia que existe entre ellas [12].

2.2.2. Características espectrales

Estas características describen la señal del habla en el dominio de la frecuencia como son los armónicos y los formantes. Los armónicos son múltiplos de la frecuencia fundamental y se reconocen por su frecuencia y amplitud [28], mientras que los formantes son amplificaciones de ciertas frecuencias en el espectro, resultantes de la resonancia en el tracto vocal caracterizados por su frecuencia, su amplitud, y su ancho de banda. En general, los dos primeros formantes, son suficientes para eliminar la ambigüedad en las vocales [29]. Podría llegar a parecer que estas características resuelven todos los problemas de segmentación, sin embargo, éstas presentan el mismo inconveniente: sólo tiene sentido extraerlas de sonidos sonoros, es decir, de señales periódicas.

Cuando hablamos de sonidos sonoros, lo hacemos pensando en aquellos que son producidos con ayuda de las cuerdas vocales, por ejemplo el fonema /e/, por el contrario, los sonidos fricativos como el del fonema /f/, son producidos con la ayuda de los elementos coarticuladores de la cavidad oral como los dientes o la lengua. Los sonidos sonoros tienen la particularidad de poseer una señal periódica al contrario de los fricativos, en la Figura 2.2 se muestra la representación oscilográfica de la palabra ‘cinco’ donde se puede apreciar este fenómeno mientras que en la Figura 2.3 se observan los formantes de una señal periódica y los de una señal aperiódica que bien podrían confundirse con ruido ambiental o con el silencio.

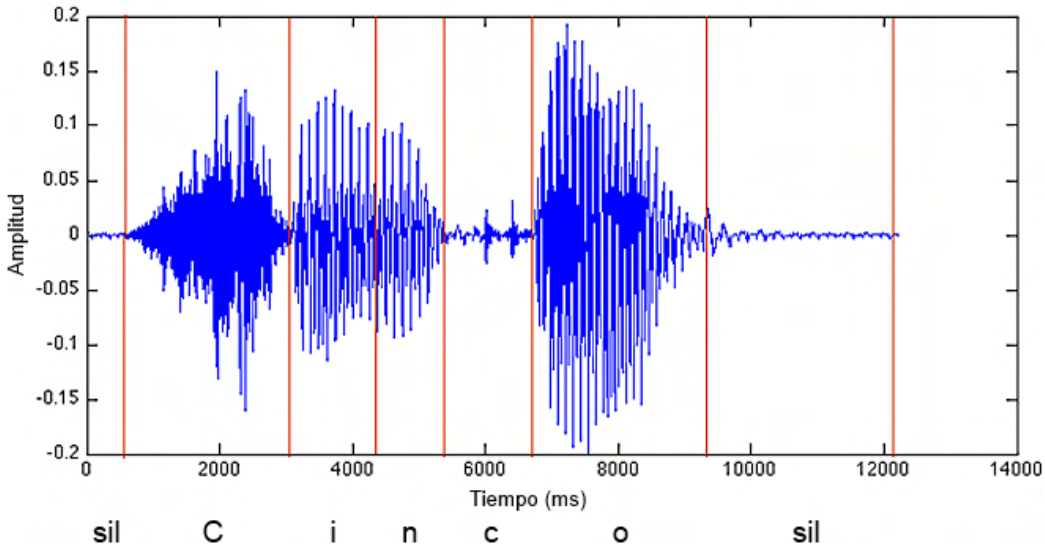

Figura 2.2: *Representación oscilográfica de la palabra 'cinco' con silencios al inicio y al final representados por 'sil'*

Existen otras características espectrales que son estándar en el reconocimiento de voz como los Coeficientes Cepstrales en las frecuencias de Mel (MFCC) o la Predicción Linear de Coeficientes Cepstrales (LPCC) [30].

2.2.3. Características de calidad en la voz

Las cualidades de la voz son estilos de habla como el neutro, entrecortado, susurrante, crujiente, duro o en falsete [31]. Se han propuesto numerosos algoritmos de filtrado automático para obtener estas características [32], algunas de ellas son el *jitter* y el *shimmer* que miden la variación ciclo a ciclo entre la duración del periodo y el pico de amplitud; en general éstos miden el ruido en la señal, el primero es un cambio indeseado y abrupto mientras que el segundo representa la refracción al pasar, por ejemplo, por un medio turbulento como cuando se habla a través de un ventilador.

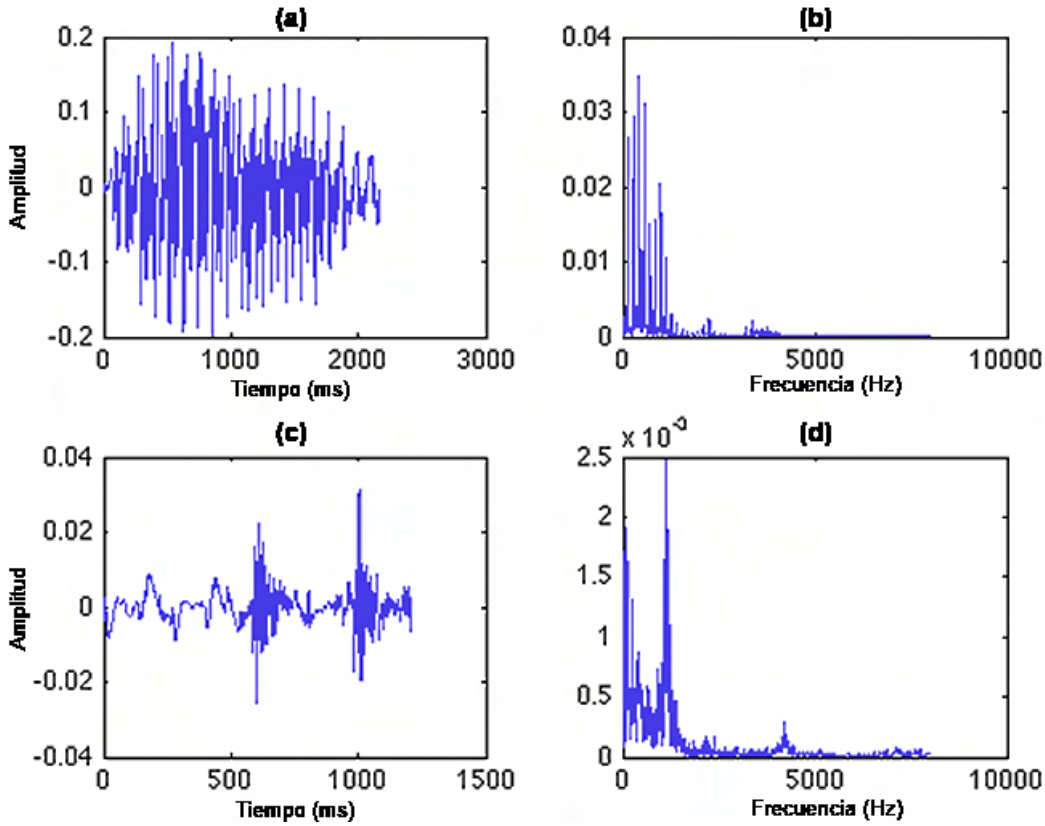

Figura 2.3: *(a) Señal periódica, (b) Formantes de la señal periódica, (c) Señal aperiódica, (d) Formantes de la señal aperiódica*

Dada la gran diversidad de características que se pueden observar en el Anexo B, no está claro exactamente cuáles son las mejores a utilizar, por esto ha surgido el enfoque de 'fuerza bruta' donde la idea es extraer un gran número de características para ir haciendo las pruebas de manera exhaustiva, un ejemplo de esto se dá en [33] donde se propone la extracción de muchas de ellas que forman un superconjunto de pruebas.

Como se mostró en la Figura 2.1, la última etapa en un sistema de reconocimiento de emociones en el habla es la de clasificación, algunos de los trabajos más representativos, se consideran a continuación.

2.3. Clasificadores

Se conoce un gran número de clasificadores automáticos que han sido aplicados en el reconocimiento de emociones, un ejemplo de clasificador lineal es el análisis discriminante lineal (LDA) [28, 34], otro tipo de clasificadores muy populares aunque más complejos son las redes neuronales artificiales (ANN) [28, 35] y las máquinas de soporte vectorial (SVM) [36, 37], también existen técnicas de clasificación que hacen uso de árboles de clasificación y regresión (CART) [38] así como otros que son especialmente buenos para clasificar las características a nivel de ventanas como lo son los modelos de mezcla de Gaussianas (GMM) [39].

Desde el Anexo A podemos extraer los porcentajes de desempeño para estos clasificadores aunque éstos varían mucho pues los estudios fueron realizados con distintas bases de datos e incluso con distintas emociones, por ejemplo: para el caso de LDA se tienen resultados que van del 50 % al 88 %, para las ANN los porcentajes están entre 50 % y 90 %, las SVM presentan un desempeños de entre el 44 % y el 92 %, los CART reportan hasta un 97 % mientras que los resultados para GMM van desde el 75 % hasta el 92 %.

Los diferentes resultados en los estudios citados anteriormente muestran que no hay un mejor clasificador. Qué clasificador arrojará los mejores resultados dependerá de los datos utilizados y de la experiencia del usuario para la elección correcta de los parámetros. En el Capítulo 4 se presentan las DBNs y RBMs como método de clasificación que comparamos con otros clasificadores como MLP, K-vecinos más cercanos (KNN), Árboles de Decisión (DT) y SVM.

Es importante mencionar que aunque se cuenta con numerosos estudios de clasificación de emociones en la voz, existen muy pocos que utilizan las arquitecturas profundas y menos aún los que las combinan con RBMs como

lo hicieron André Stuhlsatz et al., en el 2011 [40]. Otra área de investigación estrechamente relacionada con las emociones en la voz es el reconocimiento de emociones en la música donde también encontramos un solo trabajo que combinara el aprendizaje profundo y las Máquinas Restringidas de Boltzmann publicado por Erik M. Schmidt y Youngmoo E. Kim también en el 2011 [41].

CAPÍTULO 3

EXTRACCIÓN DE CARACTERÍSTICAS

La extracción de características de la voz para el reconocimiento de emociones resulta de gran importancia ya que son estas características las que serán examinadas para determinar la clase a la que pertenecerá el sonido.

Para extraer estas características, primero se deberá pensar en las que resulten más adecuadas para resolver el problema, en particular el de emociones, ya que podría no ser suficiente extraer el tono y la frecuencia. Incluso se debe considerar la duración del intervalo que se analizará para determinar si dichas características serán de utilidad en la clasificación; tomando en cuenta la duración de este intervalo, pueden ser divididas en dos: las locales, que son extraídas de secciones o ventanas de la señal y las globales que son obtenidas mediante cálculos estadísticos de los elementos locales.

Diversos estudios han mostrado que el uso de características globales tienden a producir mejores resultados en lo que concierne a la exactitud y tiempo de clasificación [26], una de las razones es que el número de características globales es mucho menor en comparación con las características locales y por

ende el tiempo requerido por las técnicas de validación es menor. Así mismo muestran que si bien las características globales ayudan a producir mejores clasificaciones, éstas sólo son eficientes para distinguir entre emociones de alta y baja energía como en el caso de la ira y la tristeza.

De igual manera se ha visto que la dimensionalidad de los vectores de entrenamiento que serán obtenidos mediante el análisis de características globales será menor, lo que dificultará el uso de clasificadores como las SVM y los Modelos Ocultos de Markov (HMM). Para estos tipos de clasificadores que funcionan mejor con vectores de alta dimensionalidad, será preferible utilizar las características extraídas mediante el análisis local [26].

Esto lleva a cuestionarnos el tamaño de la ventana. Una opción a seguir es segmentar la señal de la voz en los fonemas que la conforman y otra es la de segmentarla por cada frase, con esto, el tamaño de la ventana y el tipo de análisis, local o global, influirán en el número de elementos de los vectores de entrenamiento. Tener muchos o pocos de éstos no implica necesariamente que se clasificarán correctamente las señales de voz ya que también son importantes las características representadas en ellos, es por esto que debemos cerciorarnos que éstas describan correctamente el contenido emocional de la voz.

Como se puede ver en la Figura 3.1, las características de la voz pueden ser agrupadas en tres categorías: continuas, espectrales y cualitativas [26]. Se ha estipulado que elementos tales como el tono y la energía transmiten la mayor parte del contenido emocional de un enunciado, esta propiedad permite diferenciar aquellas emociones que tienen una alta activación de las que poseen una baja [42], las características más usadas en esta categoría son: la frecuencia fundamental, la energía, la duración y los formantes.

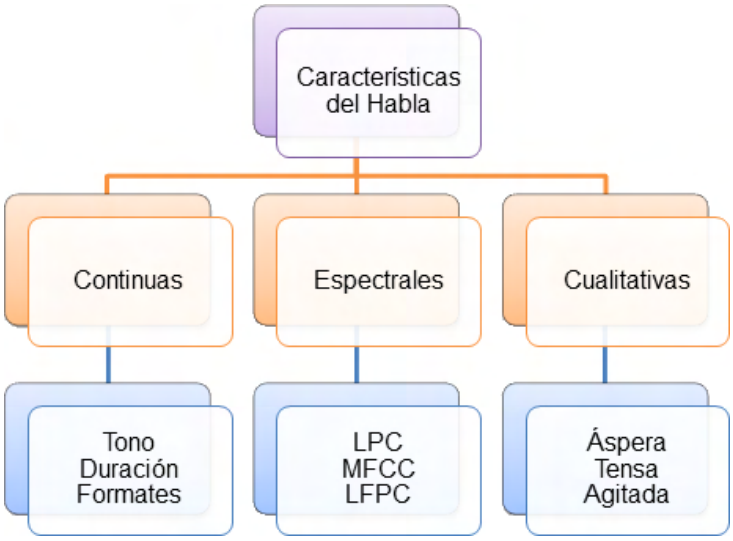

Figura 3.1: *Características del habla agrupadas en tres categorías: continuas, espectrales y cualitativas*

Se encuentran elementos como el volumen, el tono y la estructura temporal [12], que intentan definir las características cualitativas. Se sabe relativamente poco de ellos pues a pesar de haber sido estudiados se utilizan términos subjetivos para describirlos como por ejemplo: áspero, tenso o agitado. Esto hace difícil su caracterización y, por consiguiente, dificulta la clasificación automática de la voz ya que una voz tensa o agitada puede asociarse a enojo, alegría y miedo de la misma forma que una voz relajada o resonante podría asociarse con tristeza.

Las características espectrales podrán extraerse mediante distintas técnicas entre las que destacan: codificación lineal predictiva (LPC), coeficientes cepstrales en las frecuencias de Mel (MFCC) y coeficientes de potencia de la frecuencia logarítmica (LFPC). Estos elementos resultan importantes debido a que 'el contenido emocional de una frase tiene impacto en la distribución de la energía espectral a lo largo del rango de frecuencias de la voz' [26].

	Enojo	Felicidad	Tristeza	Miedo	Disgusto
Velocidad	Más rápida	Rápida o lenta	Más lenta	Muy rápida	Mucho más rápida
Tono promedio	Muy alto	Más alto	Más bajo	Muy alto	Muy bajo
Intensidad	Alta	Alta	Baja	Normal	Baja
Calidad	Agitada	A todo volumen, jadeante	Resonante, resoplante	Irregular	Quejumbrosa
Cambios en el tono	Abruptos	Suaves, hacia arriba	Suaves, hacia abajo	Normales	Amplios

Tabla 3.1: *Emociones y parámetros de la voz tomado de Cowie et al. [12], 2001*

Algo más que sabemos sobre la energía en la voz es que la frecuencia fundamental y sus armónicos cambian bajo distintos estados emocionales, bajo esta premisa Teager expone que 'escuchar es el proceso de detectar energía' [43].

La velocidad con que vibran las cuerdas vocales determina la frecuencia fundamental de la señal de la voz mientras que los patrones en el movimiento de la F_0 describen la entonación, por ejemplo, al realizar una pregunta. Como hemos visto, el tono, el acento y el ritmo son elementos que estudia la prosodia, es desde este aspecto fonético-acústico que de la Tabla 3.1 se desprende lo siguiente:

- El habla se puede relacionar con las emociones arquetípicas; enojo, felicidad, tristeza, miedo y disgusto. Las medidas del habla que parecen ser indicadores confiables son las medidas acústicas continuas, particularmente las relacionadas con el tono, intensidad y duración.

- El conocimiento que tenemos sobre las emociones es subjetivo, irregular e inconcluso dado que hay contradicciones en las emociones arquetípicas. Por ejemplo, algunos autores reportan diversas duraciones y velocidades en la expresión de la ira, la felicidad y el miedo [12]. Aunado a esto, el conocimiento que tenemos sobre la calidad de la voz también es incompleto, se mencionan en múltiples ocasiones los atributos de la calidad de la voz pero son, en su mayoría, subjetivos y juzgados solo de 'oído'.

- La falta de integración entre la paralingüística y la lingüística, es decir, no se han logrado conjuntar los atributos de una y otra. La paralingüística es parte del estudio de la comunicación humana que investiga los elementos que acompañan las emisiones propiamente lingüísticas y que constituyen señales e indicios normalmente no verbales. Además, podemos inferir a partir de ella que los atributos paralingüístcos se relacionan con las emociones arquetípicas y los atributos lingüísticos están ligados a las emociones no arquetípicas.

- Algunos atributos de la voz parecen estar asociados con características generales de las emociones, por ejemplo la activación; una activación positiva está relacionada con un alto nivel en la media y rango de F_0, tal es el caso de la felicidad, el miedo, la ira y la sorpresa mientras que una activación negativa está relacionada con una baja en la media y rango de F_0 por ejemplo en la tristeza y el aburrimiento.

Como se ha venido exponiendo en esta sección, seleccionar los atributos que describan correctamente una emoción es una tarea de importancia. Es para esta selección que se han explorado distintas opciones como es el caso del volumen en la voz pues suele ser un indicador intuitivo de la emoción, algunos autores [43] la miden como una función directa del voltaje del micrófono

aunque normalmente depende de la distancia entre éste y el hablante, de la dirección y del ambiente. Para abordar este problema, se ha propuesto prestar atención al estrés del hablante ya que esto ayuda a distinguir el volumen [44]. Otra opción es utilizar una distribución de energía ya que los brincos hacia las vocales y en contra de las consonantes son indicador de un alto volumen.

La duración de las pausas entre las palabras, los fonemas y los atributos parece ser otro buen indicador de las emociones ya que la velocidad del habla se puede medir en palabras por minuto, esto nos lleva a la necesidad de detectar los límites entre las palabras lo que es una tarea complicada. Para aminorar este problema se proponen otras formas de medir la velocidad del habla que funcionan mejor para la extracción automática, tal es el caso de la detección de fonemas mediante el análisis del núcleo de la sílaba o encontrar los límites entre las características del habla como el de la 'explosión fricativa'.

Habiendo descrito de manera general algunas características que pueden ser extraídas de la señal del habla, se manifiesta la importancia de los atributos a los que nombramos 'relevantes' pues en base a la experimentación que realizamos, creemos que ayudan a caracterizar de manera eficaz las emociones que se presentan en la voz.

A continuación se plantean brevemente los atributos utilizados en este trabajo. Una revisión más detallada sobre la extracción de éstos, se puede ver el Anexo C.

3.1. Tono

El sonido se produce cuando el aire que sale a presión de los pulmones pasa por las cuerdas vocales y la cavidad oral, la velocidad a la que estas cuerdas vibran determina la frecuencia fundamental o tono de la voz, por

lo que las emociones humanas tienen un fuerte efecto sobre los contornos de esta característica. Por este motivo los cambios en el contorno de F_0 se han utilizado ampliamente.

3.2. Coeficientes cepstrales en las frecuencias de Mel

Ya que el oído humano percibe el sonido en una escala logarítmica [45], es necesario utilizar una transformación de las frecuencias a la escala perceptual. Un método para realizar esta transformación es hacer uso de la escala de Mel que es una aproximación a la escala perceptual humana; el punto de referencia entre ésta y la medición normal de la frecuencia se define mediante la asignación de un campo perceptivo de 1000 Mels a un tono de 1000 Hz, 40 dB por encima del umbral del oyente y, a partir de los 500 Hz se definen intervalos cada vez más grandes para producir incrementos iguales de tono, como resultado, cuatro octavas en la escala de hercios son alrededor de dos octavas en la escala Mel.

Los MFCCs son las amplitudes del espectro resultante de tomar la transformada de Fourier de la ventana de una señal, mapear la energía del espectro obtenido a la escala Mel usando una función ventana triangular, calcular el logaritmo de la energía de cada frecuencia Mel y aplicar la transformada de coseno discreta. En el Apéndice C se da una descripción más amplia.

3.3. Tasa de cruces por cero

La Tasa de Cruces por Cero (ZCR) proporciona una idea general de la distribución en frecuencia de la señal, pues mide las veces que la señal de voz

pasa por el nivel cero. Tener un valor de ZCR elevado indica que el segmento de voz tiene un contenido espectral importante en alta frecuencia, mientras que una tasa baja implica que casi toda la señal está en baja frecuencia. Esta distinción del espectro permite obtener otra forma de separar los segmentos de voz sonoros, de los sordos; un segmento sonoro posee un espectro centrado en baja frecuencia y uno sordo tiene una componente superior en alta frecuencia. El mayor inconveniente de la ZCR es que se ve muy influenciada por el ruido de fondo, ya que los segmentos no hablados suelen contener frecuencias más altas que los hablados. Recordemos que el rango de frecuencias en la voz va desde los 85 Hz a los 255 Hz en el habla típica, mientras que para las cuerdas vocales entrenadas va desde los 80 Hz para los bajos hasta los 1100 Hz para los sopranos.

3.4. Energía

Por naturaleza, la energía asociada con el habla es variable en el tiempo. De ahí el interés por saber cómo ésta característica está cambiando, más específicamente, cómo lo está haciendo en una región del segmento hablado, como en el caso de una palabra. Como sabemos, la señal de voz consiste en segmentos hablados y no hablados, es en estas regiones que la energía asociada a la presencia de la voz es grande en comparación con la región sin voz. Por lo tanto, la energía en segmentos cortos se puede utilizar para determinar aquellas regiones que nos sean de interés.

CAPÍTULO 4

APRENDIZAJE PROFUNDO

Algunos autores sugieren que las arquitecturas profundas son mucho más eficientes, en términos de los elementos computacionales requeridos, que las arquitecturas simples o de poca profundidad, incluyendo los modelos ocultos de Markov, las redes neuronales con una sola capa oculta y las máquinas de soporte vectorial entre otras [46]. Por desgracia, este tipo de redes profundas son muy difíciles de entrenar, para superar este problema en el 2006 se introdujo un algoritmo de aprendizaje no supervisado que hizo posible dicho entrenamiento [2]. En las siguientes secciones se dará una breve introducción a estas arquitecturas profundas y a las RBMs al mismo tiempo que se revisarán sus algoritmos de entrenamiento.

4.1. Introducción

Una motivación clave para el aprendizaje profundo o deep learning (DL) es la biología, ésta nos dice que el cerebro funciona de forma 'profunda', dicha naturaleza, a la que llamaremos jerárquica, proviene de la observación de que

las capas superiores representan conceptos cada vez más abstractos, es por esto que se cree que, esta estructura jerárquica empleada por el Neocórtex, es la respuesta a gran parte de su poder [47, 48, 49].

El aprendizaje profundo es un paradigma del aprendizaje maquinal que se centra en el uso de los modelos jerárquicos de datos, éste plantea la hipótesis de que con el fin de aprender las representaciones de alto nivel de los datos, se necesita una jerarquía de representaciones intermedias. Por ejemplo, en el caso de la visión computacional, el primer nivel de representación podría ser la obtención de los pixeles, el segundo nivel podría reconocer líneas y contornos, mientras que las representaciones de nivel superior podrían reconocer partes y objetos como se puede ver en la Figura 4.1. La hipótesis es que, si se permite que la red encuentre representaciones en varios niveles de abstracción, se obtendrán mejores resultados, pues cada capa irá encontrando patrones en las capas más bajas y representando conceptos más abstractos en las superiores. Aunque parece ser una buena idea, en la práctica no resulta tan simple como apilar muchas capas, no obstante, los recientes avances en los algoritmos de aprendizaje para arquitecturas profundas, han hecho posible que estos sistemas sean factibles [2].

En este trabajo se propone hacer uso de una arquitectura profunda que utilice como bloque de construcción las máquinas restringidas de Boltzmann (RBMs) que pueden ser apiladas para obtener distintos niveles de especialización [50]. Ésto en lugar de las redes neuronales artificiales (ANN) empleadas tradicionalmente, sin embargo, se puede ejemplificar esta arquitectura profunda recurriendo a ellas; teniendo como base una ANN en la primer capa y deseando extenderla a una nueva, se plantea utilizar la salida de esta primera como entrada para la siguiente. Este concepto se irá explicando a lo largo de las siguientes secciones aunque, si se desea, se puede ver representado en la Figura 4.6.

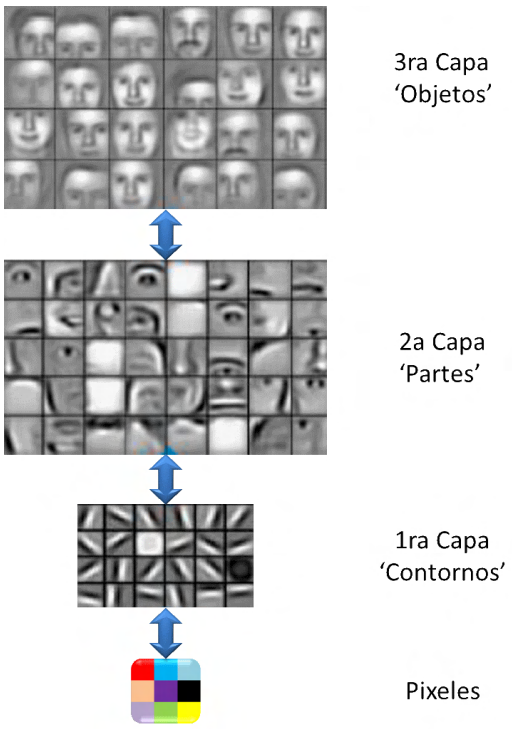

Figura 4.1: *Niveles de representación jerárquica en la visión por computadora*

4.2. Máquinas Restringidas de Boltzmann

La máquina de Boltzmann (BM), representada en la Figura 4.2, fue desarrollada por Geoffrey Hinton y Terry Sejnowski en 1983 [51]. Ésta es un tipo de red neuronal donde todas las neuronas están conectadas entre si y tiene la particularidad de que toma decisiones estocásticas sobre si una neurona estará activada o no, es decir, son construidas introduciendo variaciones probabilistas a los pesos de la red. A esta máquina se le presenta un conjunto de vectores de entrenamiento que deberá aprender a clasificar con alta probabilidad, para lograrlo la BM debe encontrar los pesos de las conexiones que logren que los vectores de datos con los que fue entrenada presenten un 'costo' o valor bajo en relación con otros ejemplos.

No obstante, las BM sin restricciones de conectividad no han demostrado ser útiles para resolver los problemas que se dan en la práctica ya que, como es de esperarse, el proceso de aprendizaje es lento en redes de gran tamaño debido a la forma en la que están construidas como se puede ver en la Figura 4.2. Es por esto que se propusieron las máquinas *restringidas* de Boltzmann (RBM), esta 'restricción' se basa en reducir el número de conexiones impidiendo que las neuronas o unidades de la misma capa se 'vean', como en la Figura 4.3.

En ambas figuras, v son las unidades visibles o capa de entrada, h las unidades ocultas o capa de salida, W las conexiones o pesos entre v y h, J son las conexiones o pesos entre las unidades h_i y h_j y L son las conexiones o pesos entre las unidades v_l y v_l.

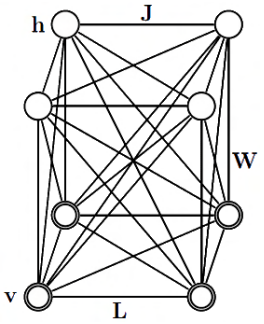

Figura 4.2: *Máquina de Boltzmann*

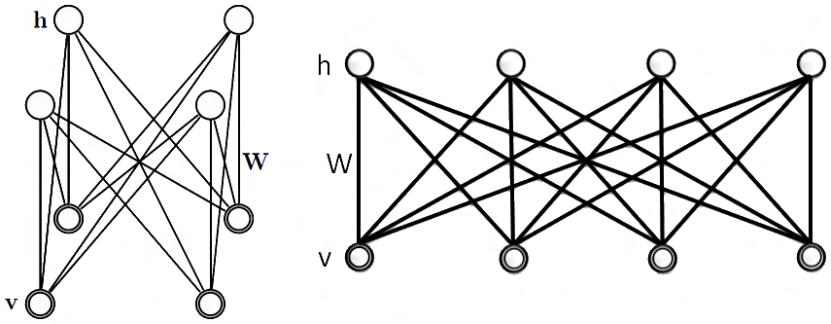

Figura 4.3: *Máquina Restringida de Boltzmann*

Sean v_i y h_j los estados de la unidad visible i y la unidad oculta j y a_i y b_j sus respectivos sesgos con los pesos $w_{i,j}$ entre v_i y h_j, la función de energía de las RBMs está dada por [52]:

$$E(v,h) = -\sum_{i \in visible} a_i b_i - \sum_{j \in oculta} b_j h_j - \sum_{i,j} v_i h_i w_{i,j} \tag{4.1}$$

O de manera simplificada:

$$E(v,h) = -a'v - b'h - h'Wv \tag{4.2}$$

La red asigna una probabilidad a cada par entre una unidad visible y un vector de unidades ocultas a través de esta función:

$$p(v,h) = \frac{\exp^{-E(v,h)}}{Z} \tag{4.3}$$

Donde la función de partición Z, está dada por la suma de todos los pares de vectores visibles y ocultos:

$$Z = \sum_{v,h} \exp^{-E(v,h)} \tag{4.4}$$

La energía libre de la entrada, es decir, de las unidades visibles, es la que se debe modificar para aumentar o reducir las probabilidades como se explica en las siguientes tres ecuaciones [3]:

$$FE(v) = -\sum_i a_i v_i - \sum_i log - \sum_{h_i} \exp^{h_i W_i x} \tag{4.5}$$

La probabilidad de que la red clasifique a un vector de unidades visibles v, está dada por sumatoria de todos los vectores ocultos:

$$p(v) = \frac{1}{Z} \sum_h \exp^{-E(v,h)} \tag{4.6}$$

Esta probabilidad se puede elevar mediante el ajuste de los pesos para reducir la energía de ese vector y así aumentar la de los otros, en la siguiente ecuación con ϵ siendo la tasa de aprendizaje, se muestra esa modificación de los pesos.

$$\Delta w_{i,j} = \epsilon(\langle v_i h_j \rangle_{datos} - \langle v_i h_j \rangle_{modelo}) \tag{4.7}$$

El gradiente de la probabilidad logarítmica de un vector de entrenamiento con respecto a un peso donde los paréntesis angulares, $\langle\rangle$, denotan la distribución de los *datos* y del *modelo* respectivamente es:

$$\frac{\partial \log p(v)}{\partial w_{i,j}} = \langle v_i h_j \rangle_{datos} - \langle v_i h_j \rangle_{modelo} \tag{4.8}$$

Debido a la estructura específica de estas redes, las unidades visibles y ocultas son condicionalmente independientes [53]:

$$p(v|h) = \prod_i p(v_i|h)$$
$$p(h|v) = \prod_j p(h_j|v) \tag{4.9}$$

Usando esta propiedad, podemos escribir:

$$p(v_j = 1|h) = \sigma(a_i + \sum_j h_j w_{i,j})$$
$$p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{i,j}) \tag{4.10}$$

Donde σ es la función sigmoidea:

$$\sigma(x) = \frac{1}{1 + \exp^{-x}} \tag{4.11}$$

Este modelo cuenta con dos fases; la fase positiva disminuye la energía de los datos de entrenamiento y la fase negativa aumenta la energía de todos los demás estados visibles que el modelo puede generar. La fase positiva es fácil de calcular debido a la Ecuación (4.9), por el contrario, la fase negativa no es fácil de calcular ya que implica sumar todos los estados posibles del modelo, por este motivo en lugar de calcular la fase negativa exacta se realizan muestras del modelo.

En resumen, la idea para entrenar el modelo es hacer que éste genere datos parecidos a aquellos que le fueron presentados como de entrenamiento,

o dicho de otra manera, queremos maximizar la probabilidad logarítmica de los datos de entrenamiento o reducir al mínimo la probabilidad logarítmica negativa de estos.

4.2.1. Muestreo y divergencia contrastiva

Como dijimos, las RBMs son un tipo de red neuronal, por lo que también funcionan actualizando los estados de algunas unidades en dependencia de otras, para actualizar el estado de una unidad i se necesita hacer uso de las ecuaciones (4.7) y (4.8). Nótese que no podemos garantizar que una unidad será activada, en todo caso podemos decir que será activada con alta probabilidad.

Las muestras de $p(x)$ se pueden conseguir mediante la ejecución de una cadena de Markov hasta la convergencia utilizando el muestreo de Gibbs [54].

El muestreo de Gibbs para N variables aleatorias $S = (S_1, \ldots, S_N)$ se realiza a través de una secuencia de muestreo con N sub-pasos de la forma $S_i \sim p(S_i|S_{-i})$ donde S_{-i} contiene las $N-1$ variables aleatorias de S excluyendo S_i. Para las RBMs, como se puede ver esquematizado en la Figura 4.4, S consiste en el conjunto de unidades visibles y ocultas y ya que son condicionalmente independientes, se puede realizar el muestreo por bloques:

$$\begin{aligned} h^0 &= p(h|v^0) \\ v^1 &= p(v|v^0) \\ h^1 &= p(h|v^1) \\ &\ldots \\ v^n &= p(v|h^{n-1}) \end{aligned} \tag{4.12}$$

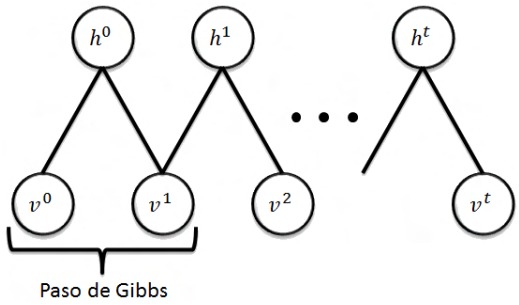

Figura 4.4: *Paso de Gibbs*

Con esta configuración, las unidades visibles se muestrean simultáneamente con los valores fijos en las unidades ocultas, de forma similar, las unidades ocultas se muestrean simultáneamente dados los valores de las unidades visibles. Un paso en la cadena de Markov se toma como sigue:

$$\begin{aligned} h^{n+1} &\sim \sigma(W'v^n + c) \\ v^{n+1} &\sim \sigma(Wh^{n+1} + b) \end{aligned} \tag{4.13}$$

Donde h^n se refiere al conjunto de todas las unidades ocultas en el paso n-ésimo de la cadena de Markov. Esto quiere decir que para el caso particular de h_i^{n+1}, se verá activada con probabilidad $\sigma(W_i'v^n + c_i)$ y, de manera similar v_j^{n+1} se verá activada con probabilidad $\sigma(W_j h^{n+1} + b_j)$. Cuando $t \to \infty$, las muestras v^t, h^t modelan correctamente a $p(v, h)$ aunque por supuesto, hacer los cálculos para $t \to \infty$ resulta computacionalmente prohibitivo. Es por esto que se propuso la Divergencia Contrastiva (CD) [55].

La CD hace uso de dos técnicas para acelerar el proceso de muestreo:

- Ya que se desea que la $p(v) \approx p_{entrenamiento}(v)$, la cadena de Markov se inicializa con un vector de entrenamiento con lo que se logra que la distribución sea cercana a p y por ende, la cadena esté próxima a converger.
- Además, la CD no espera a que la cadena converja, realiza un número k de pasos de Gibbs antes de detenerse. En la práctica $k = 1$ es suficiente [3].

Para utilizar la DBN para clasificación, un vector de entrada se le presenta a la capa visible del primer nivel, pasando hacia arriba las salidas a través de la DBN hasta que se llega a la última capa oculta. En la RBM superior se elige la unidad que tiene la menor energía libre [2], una forma más simple de usar la DBN para clasificación es simplemente añadir una última capa consistente en un clasificador estándar y entrenar todo el modelo como si se tratase de una red neuronal feedforward con backpropagation.

4.3. Ejemplo básico del funcionamiento de la RBM

El siguiente ejemplo tomado desde [56] puede ayudar a comprender mejor el funcionamiento de las RBMs. Supongamos que se le pide a un grupo de personas que, de un conjunto de películas digan cuáles les han gustado y cuáles no, con esta información la RBM descubrirá el género que prefieren.

Para este ejemplo se configuran las capas de la red como sigue:

- Una capa de unidades visibles que modelan las preferencias de los usuarios sobre las películas.
- Una capa de unidades ocultas, el género que deseamos descubrir.

De un conjunto de seis películas -Harry Potter, Avatar, Lord of The Rings 3, Gladiator, Titanic y Glitter- se les pide a los usuarios que digan qué películas les gustan, de este conjunto dos géneros se pueden identificar: ciencia ficción -Harry Potter, Avatar, Lord of The Rings 3- y ganadoras del Oscar -Gladiator, Titanic y Glitter- con esta información, la estructura de la RBM quedaría como sigue:

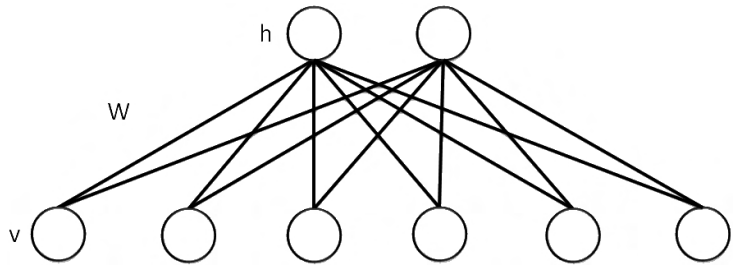

Figura 4.5: *RBM de ejemplo con seis unidades visibles -películas- y dos unidades ocultas -géneros-*

Las dos unidades ocultas corresponden a los géneros ciencia ficción y ganadoras del Oscar y las seis visibles a las preferencias que las personas han elegido sobre el conjunto de películas, con lo que tenemos un vector binario de 6 posiciones donde los valores 1 y 0 corresponden a aceptación y desaprobación respectivamente. Véase la Tabla 4.1.

Entonces, el primer paso sería presentar un vector de entrenamiento a la red y realizar tantos pasos de Gibbs sean necesarios para converger, en la práctica, como hemos dicho, un solo paso suele ser suficiente. Posteriormente se repite el proceso para todos los vectores de entrenamiento, usualmente se seleccionan particiones de los datos para entrenamiento, prueba y validación aunque aquí, por simplicidad, todos los vectores son utilizados para el entrenamiento.

[Harry Potter, Avatar, LOTR 3 Gladiator, Titanic, Glitter]	
Vector de entrenamiento	Etiqueta/clase
[1,1,1,0,0,0]	Ciencia ficción
[1,0,1,0,0,0]	Ciencia ficción
[1,1,1,0,0,0]	Ciencia ficción
[0,0,1,1,1,0]	Ganadora del Óscar
[0,0,1,0,1,0]	Ganadora del Óscar
[0,0,1,1,1,0]	Ganadora del Óscar

Tabla 4.1: *Ejemplo de un conjunto de entrenamiento para la RBM, los vectores binarios representan la selección de películas que prefiere cada persona*

El proceso para la modificación de los pesos utilizando Gibbs se debe realizar para cada ejemplo de entrenamiento y es el siguiente:

- Tomar un vector de entrenamiento y presentarlo a las unidades visibles.
- Actualizar el estado de las unidades ocultas usando la función sigmoide, y para cada arco $e_{i,j}$ calcular $a(e_{i,j}) = v_i * h_j$ donde la a denota el valor de activación, lo que en la Ecuación (4.7) correspondería a $\langle v_i h_j \rangle_{datos}$
- Reconstruir las unidades visibles de forma similar: para cada unidad visible i calcular a_i y actualizar su estado.
- Actualizar de nuevo las unidades ocultas y calcular $a(e_{i,j}) = v_i * h_j$ para cada arco, que correspondería a $\langle v_i h_j \rangle_{modelo}$
- Actualizar los pesos de cada arco $e_{i,j}$ mediante la Ecuación (4.7).

Tras actualizar los pesos de la red se obtuvieron los siguientes valores de activación en las unidades ocultas, de ellos que se eligieron los más altos:

	Unidad oculta 1	Unidad oculta 2
HarryPotter	-7.08986885	4.96606654
Avatar	-5.18354129	2.27197472
LOTR3	2.51720193	4.11061383
Gladiator	6.74833901	-4.00505343
Titanic	3.25474524	-5.59606865
Glitter	-2.81563804	-2.91540988

Tabla 4.2: *Activaciones de las unidades ocultas de la RBM*

Estas unidades que fueron seleccionadas representan la clase a la que pertenece el ejemplo presentado a la red, con estos resultados podemos decir que la clasificación ha sido satisfactoria. Además se confirma lo que habíamos supuesto; la primera unidad oculta corresponde a los ganadores del Oscar y la segunda a ciencia ficción.

4.4. Redes de creencia profunda

Una Red de Creencia Profunda (DBN) es un tipo red neuronal con una arquitectura profunda, es decir, con muchas capas ocultas. Se compone de una capa de entrada que contiene las unidades visibles, un número ℓ de capas ocultas y finalmente una capa de salida que tiene una unidad para cada clase a clasificar. En la Figura 4.6 se muestra una DBN con $\ell = 3$ a la que se le puede agregar una capa que funcione como clasificador.

Las DBNs modelan la distribución entre el vector de unidades visibles v y las ℓ capas ocultas h^k de la siguiente forma:

$$P(v, h^1, \ldots, h\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k | h^{k+1}) \right) P(h^{\ell-1}, h^\ell) \tag{4.14}$$

Donde $v = h^0$ y $P(h^{k-1}|h^k)$ es una distribución para las unidades visibles condicionada a las unidades ocultas de la Máquina Restringida de Boltzmann, en el nivel k, y $P(h^{l-1}|h^l)$ es la distribución para la RBM del nivel superior. Los parámetros de una DBN son los pesos w^j entre las unidades de las capas $j-1$ y j y el sesgo b^j de la capa j.

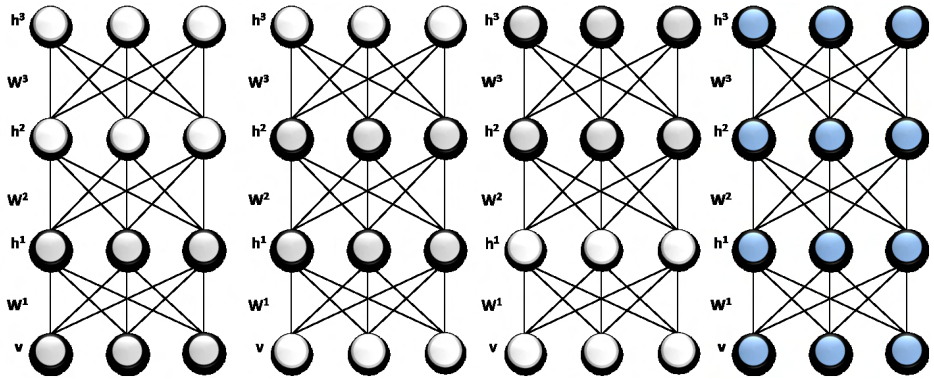

Figura 4.6: *Pasos en el entrenamiento de una DBN*

Para entrenar una DBN se sigue un algoritmo voraz no supervisado que se aplica a las RBMs que la constituyen [57], el proceso es como sigue:

Paso 1. Entrenar la primer capa como una RBM.

Paso 2. Obtener la salida de (1) y utilizarla como fuente de datos para la siguiente capa.

Paso 3. Entrenar la siguiente RBM, utilizando los datos obtenidos en (2) como vectores de entrenamiento para la capa visible.

Paso 4. Repetir los pasos (2 y 3) para cada capa.

Paso 5. Realizar un ajuste fino a los parámetros de la DBN.

Habiendo ejecutado los procedimientos enunciados en este capítulo encontramos que el entrenamiento de las RBMs en una arquitectura profunda sí es factible y, consecuentemente, la idea de los vectores donde se modelan las preferencias personales, se puede utilizar para que en lugar de preferencias, represente las características extraídas de la señal del habla donde cada elemento de ese vector represente una de ellas. Con esto podemos empezar a reconocer un procedimiento para la clasificación de emociones en el habla.

En el capítulo siguiente se revisará la base de datos de habla emocionada que utilizaremos para el entrenamiento del sistema propuesto en este trabajo.

CAPÍTULO 5

BASE DE DATOS

Como podemos ver en el Anexo A, se han encontrado diversas bases de datos que han sido utilizadas para el reconocimiento de emociones siendo una sola de ellas, Friedrich-Alexander University- AIBO [58], del tipo 'espontaneo' donde las emociones no se han actuado, el resto son predominantemente aquellas donde se utilizan actores.

Algunos inconvenientes que existen con estos conjuntos de datos se enlistan a continuación:

- La mayoría de las bases de datos disponibles públicamente contienen un número reducido de frases, aproximadamente 500-1000, habladas por un número reducido de locutores.
- Debido a la naturaleza de los datos espontáneos, las emociones no están equilibradas.
- Como cada una de éstas se crea con un propósito específico en mente, los entornos de grabación son diferentes para cada situación.
- No hay una manera estándar para elegir la calidad y cantidad de las personas encargadas de etiquetar las frases. Por ejemplo, para la base

de datos FAU Aibo, 5 expertos etiquetan los datos, mientras que para la de habla Danesa, lo hicieron 20 alumnos.

- Debido al punto anterior, el número de emociones también es diferente y, como son inconsistentes, realizar pruebas entre ellas no es sencillo.

Es por esto que en este trabajo se ha utilizado INTERFACE [22], que es una base de datos creada en el Center for Language and Speech Technologies and Applications (TALP), de la Universidad Politécnica de Catalunya (UPC), con el propósito de estudiar el habla con emociones y la síntesis de voz. Las frases aquí contenidas expresan seis emociones más cinco variaciones de neutral, en cuatro idiomas de los cuales se utiliza el español junto con las transcripciones de las frases habladas. Cabe mencionar que los hablantes son actores profesionales, un hombre y una mujer.

5.1. Estructura

Las emociones que se muestran a continuación son las más utilizadas en el análisis y síntesis del habla emocionada; además, los estilos neutros también fueron definidos como una referencia a la expresión emocional, estas variaciones en el estilo son: lento, suave, fuerte y rápido como se muestra en la Tabla 5.1.

5.2. Corpus

El corpus consiste en 184 oraciones incluyendo palabras aisladas, oraciones y el extracto de un texto en un contexto emocional neutral, también se han incluido las formas afirmativas e interrogativas de esas mismas oraciones. La distribución se puede ver en la Tabla 5.2 y en la 5.3 se muestra un extracto de las oraciones.

Español	
6 emociones	A = enojo D = disgusto F = miedo J = alegría S = sorpresa T = tristeza
Variaciones de 'Neutral'	H = neutral/fuerte L = neutral/suave N = neutral/normal W = neutral/lento Z = neutral/rápido

Tabla 5.1: *Lista de emociones y estilos de habla*

Identificador (yyy)	Tipo de oración
001 - 100	Oraciones afirmativas
101 - 134	Oraciones interrogativas
135 - 150	Párrafos
151 - 160	Dígitos y números
161 - 184	Palabras aisladas

Tabla 5.2: *Lista de emociones y estilos de habla*

5.3. Evaluación subjetiva

Los creadores de esta base ponen a nuestra disposición el estudio que realizaron para evaluarla y que se muestra a continuación [22]. Dicha evaluación consistió en realizar pruebas subjetivas donde participaron 16 estudiantes de

Emoción	Tipo	Transcripción
Enojo	afirmativo	el presidente de la federación portuguesa de fútbol
Enojo	interrogativo	entraña riesgos la genética sintética
Tristeza	afirmativo	abría sus puertas a estos flacos alumnos afroamericanos
Tristeza	interrogativo	cómo se supone que vamos a hacerlo
Alegría	afirmativo	cuando todavía eran baratos el vodka y el caviar
Alegría	interrogativo	has comido todo lo que te han puesto en el plato
Miedo	afirmativo	la tensión volvió a aumentar el domingo
Miedo	interrogativo	y dices que tienes la colección completa
Disgusto	afirmativo	fue inyectado en el abdomen y en una pierna
Disgusto	interrogativo	de cuántos estamos hablando exactamente
Sorpresa	afirmativo	sino el país mental de un patriota enloquecido
Sorpresa	interrogativo	desde cuándo dices que come así

Tabla 5.3: *Extracto de oraciones transcritas de la base de datos*

ingeniería de la UPC como oyentes no profesionales, en estas pruebas fueron reproducidas 56 oraciones, ocho de ellas por cada una de las siete emociones. Estos datos son de mucha utilidad pues contra ellos podremos comparar los resultados obtenidos en esta propuesta.

Cada oyente decidió qué emoción correspondía a cada expresión y la intensidad percibida en una escala de uno a cinco. Una segunda opción podía ser seleccionada en el caso de que la primera no fuera clara, además, para evitar que los oyentes tuvieran una referencia inmediata, las grabaciones fueron alternadas entre el hablante femenino y el masculino.

Los resultados de esta prueba subjetiva muestran que más del 80 % de las frases fueron clasificadas correctamente con la primera elección y, de considerarse la segunda elección, más del 90 %, esto se constata en las Ta-

blas 5.4 y 5.5. Cabe mencionar que cada expresión fue correctamente clasificada por al menos la mitad de los oyentes y que los errores fueron cometidos en las palabras o frases cortas, mientras que todas las oraciones y textos largos fueron clasificadas acertadamente en el primer intento por todos los oyentes.

	S	J	A	F	D	T	N	
S	89	20	7	0	6	2	4	128
J	0	115	7	0	2	2	2	128
A	2	14	85	2	5	5	15	128
F	4	1	1	103	5	13	1	128
T	2	1	2	5	106	3	9	128
D	1	3	1	16	3	101	3	128
N	0	2	2	1	4	1	118	128
	98	156	105	127	131	127	152	896

Tabla 5.4: *Resultados de la prueba subjetiva tomados de [22]. Los valores en las columnas representan el número de oraciones reconocidas contra las emociones reales en cada fila. A=enojo, D=disgusto, F=miedo, J=alegría, S=sorpresa, T=tristeza y N=neutral*

Tasa de error			
S	30.47 %	T	21.09 %
J	10.16 %	D	17.19 %
A	33.59 %	N	7.81 %
F	19.53 %		

Tabla 5.5: *Tasas de error en la prueba subjetiva tomadas de [22]. A=enojo, D=disgusto, F=miedo, J=alegría, S=sorpresa, T=tristeza y N=neutral*

5.4. Evaluación automática

Otra evaluación que los autores de esta base de datos ponen a nuestra disposición está basada en los modelos ocultos de Markov [59]. Para realizar esta experimentación se eligieron 100 frases afirmativas pronunciadas en las siete emociones; enojo, disgusto, miedo, alegría, sorpresa, tristeza y neutral. Habladas en dos sesiones por ambos actores, el masculino y el femenino.

Esto lleva a un total de $2*2*7*100 = 2,800$ declaraciones que se dividieron en dos grupos no superpuestos: uno con $2,227$ declaraciones con fines de entrenamiento, y otro con 555 expresiones de prueba. Ambos conjuntos fueron diseñados de tal manera que el contenido de cada emoción, sesión, hablante y oración estén equilibrados a través de los grupos de entrenamiento y prueba.

De estas frases se extrajeron la primera y segunda derivada del logaritmo de la energía media y el contorno silábico como características relacionadas con la energía, por otro lado, como características relacionadas con el tono se extrajeron F_0 y sus dos primeras derivadas, el logaritmo de esas dos derivadas y el contorno silábico del tono con sus respectivas dos derivadas.

Con estas características se entrenaron siete HMMs, una para cada emoción con 1, 8, 16, 32 y 64 estados. En la etapa de reconocimiento se utilizó la máxima verosimilitud para elegir la HMM para cada frase. Los resultados se muestran en la Tabla 5.6.

En el Capítulo 8 se muestran los resultados obtenidos tras la experimentación con siete emociones. Estos resultados sugieren que nuestra propuesta, aún cuando el trabajo realizado con siete emociones es poco, se comporta de manera positiva.

Número de estados				
1	8	16	32	64
60.4 %	73.9 %	81.4 %	81.4 %	82.5 %

Tabla 5.6: *Tasas de aciertos en la prueba automática tomadas de [59].*

CAPÍTULO 6

EXPERIMENTACIÓN Y RESULTADOS

En este capítulo se mostrará las selección de parámetros utilizados en los experimentos realizados, así mismo se ofrecen los resultados obtenidos tras la aplicación de éstos y se verá la comparación de los porcentajes de error encontrados al clasificar mediante RBM y DBN con los de otros métodos que son comúnmente utilizados en la literatura, también se revisa la forma en que los datos fueron separados en los conjuntos de entrenamiento, prueba y validación de los que además se extraerán las características que permitirán dicha clasificación.

Los resultados aquí mostrados, se obtuvieron tras seleccionar las emociones neutral, alegría y tristeza como subconjunto de prueba. Con estas tres, se encontró una tasa de error del 2.51 %, que posteriormente, se comparó con las de otros clasificadores. Esto nos permitió darnos cuenta que, con los parámetros correctos, nuestro sistema se desempeña comparativamente mejor.

6.1. Conjuntos de datos

De la base de datos hemos seleccionado un conjunto reducido de tres emociones del hablante femenino con el objetivo de tener un mayor control sobre los experimentos mediante la reducción de variables, esta reducción nos permitió tener una mejor idea sobre el comportamiento del sistema. A éste subconjunto de emociones pertenecen, la neutral, la alegría y la tristeza. La primera de ellas es fundamental dado que proporciona el punto de comparación entre las otras dos, que poseen valores de activación opuestos. El número total de audios contenidos en la base de datos para estas tres emociones es de 1100, de éstos, 368 son neutrales, 366 de alegría y 366 de tristeza. En el Capítulo 8 se discute la extensión a más emociones.

Pensando en la extensión a más emociones y distintos hablantes, implementamos un mecanismo que intenta reducir el sesgo que se puede presentar cuando el número de audios por emoción está desbalanceado. Este mecanismo consiste en seleccionar aleatoria mente, dentro de la misma emoción, tantos audios hagan falta para igualar el de aquella con mayor número de ejemplos. Para el caso particular de las tres emociones, se agregaron dos a alegría y dos a tristeza, con lo que se logró que todas las emociones tuvieran 368 ejemplos y por lo tanto, que el 'tamaño del lote' con el que se entrenará la red fuera el mismo. Se debe entender por tamaño del lote como el subconjunto de de vectores de entrenamiento que le son presentados a la red en cada iteración de la divergencia contrastiva vista en la subsección 4.2.1.

El particionamiento de los datos fue seleccionado, por emoción, de la siguiente manera: 70 % para entrenamiento, 25 % para prueba y 5 % para validación. Es de estos datos de los que se extrajeron las características de las que se habla a continuación.

6.2. Selección de características

A pesar de que las características más utilizadas para el reconocimiento de emociones en la voz son los MFCCs [60], se ha discutido ampliamente el uso de otras como las prosódicas, que en conjunto, han reportado una importante mejora en la discriminación de emociones [61].

Teniendo esto en mente, propusimos el siguiente conjunto de características que fueron extraídas de las grabaciones de audio utilizando la herramienta OpenSMILE [62]. Con esto, obtuvimos un vector de 30 dimensiones.

- 12 MFCCs y su primera derivada
- Promedio de F_0 y su primera derivada
- Promedio de los cruces por cero y su primera derivada
- Energía y su primera derivada

Para obtener una primera aproximación acerca de la distribución de los patrones y la dificultad de la tarea de clasificación, hemos aplicado dos técnicas de reducción de dimensionalidad a los datos para asignarlos a un espacio de dos dimensiones, éstas fueron el Análisis de Componentes Principales (PCA), y t-SNE, del toolbox de van der Maaten [63] que se muestran en las Figuras 6.1 y 6.2.

El objetivo del PCA [64] es identificar patrones en los datos tomando en cuenta sus similitudes y diferencias. Y ya que encontrar relaciones en datos cuya dimensión es alta resulta complicado, PCA se convierte en una técnica conveniente para analizarlos. Éste es un procedimiento que convierte un conjunto de variables correlacionadas en un conjunto de variables independientes, denominadas componentes principales. Esto implica una reducción en el número de variables originales, donde el primer componente principal tiene la varianza más grande y el último la menor. En otras palabras, los

componentes principales de los datos se obtienen ordenando los eigenvalores de los eigenvectores que se encuentran con la ayuda de la matriz de covarianza de los datos. Es con esto que el número de eigenvalores ordenados que mantengamos, determinará la nueva dimensionalidad de los datos.

Por otro lado t-SNE [65], que es una técnica de reducción no lineal de dimensionalidad para la visualización de datos altamente dimensionales propuesta por el mismo autor de las RBMs, funciona convirtiendo las distancias euclidianas entre los puntos en probabilidades, donde la similitud entre x_i y x_j es la probabilidad condicional P_{ij} dada por:

$$P_{ij} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)}{\sum_{k \neq l} \exp\left(-\frac{\|x_k - x_l\|^2}{2\sigma^2}\right)} \tag{6.1}$$

Donde k es el número de vecinos y σ es la varianza, para cada x_i la $P_{ii} = 0$, esto quiere decir que mientras más se parezcan los puntos, mayor será la probabilidad P_{ij}.

Como podemos ver en las Figuras 6.1 y 6.2, todas las clases están bien distribuidas y claramente agrupadas, sin embargo, la técnica lineal PCA muestra una superposición entre alegría y tristeza. Mientras que la técnica no lineal t-SNE, muestra una distribución bidimensional un poco más complicada pero con una mejor discriminación de los datos que revela las complejas relaciones que existen entre las características. Esto sugiere que las 30 características seleccionadas representan correctamente los datos y que la tarea de clasificación es realizable. Es con estas 30 características extraídas de las oraciones que se entrenó la red con el conjunto de entrenamiento y se evaluó con el de prueba. A continuación se describen los parámetros utilizados en dicho entrenamiento y la metodología que se siguió para escogerlos.

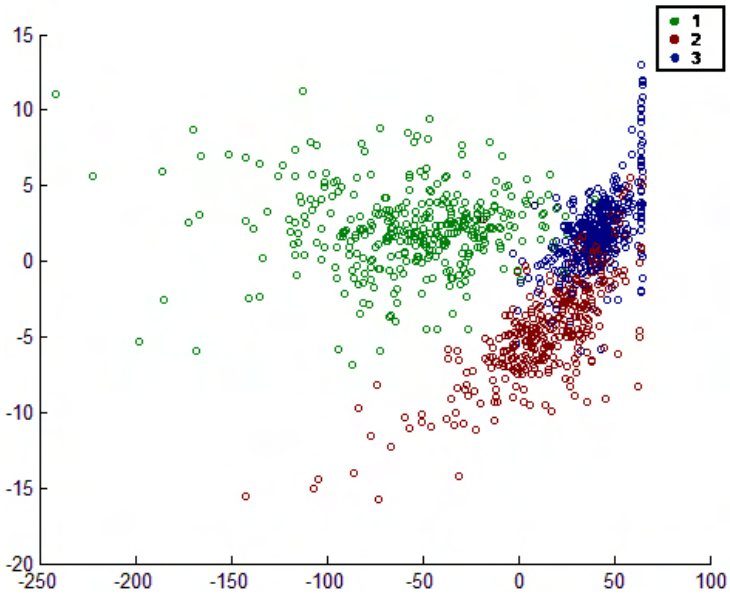

Figura 6.1: *Proyección de los vectores de características de las tres clases de emociones usando PCA: 1) neutral, 2) alegría y 3) tristeza*

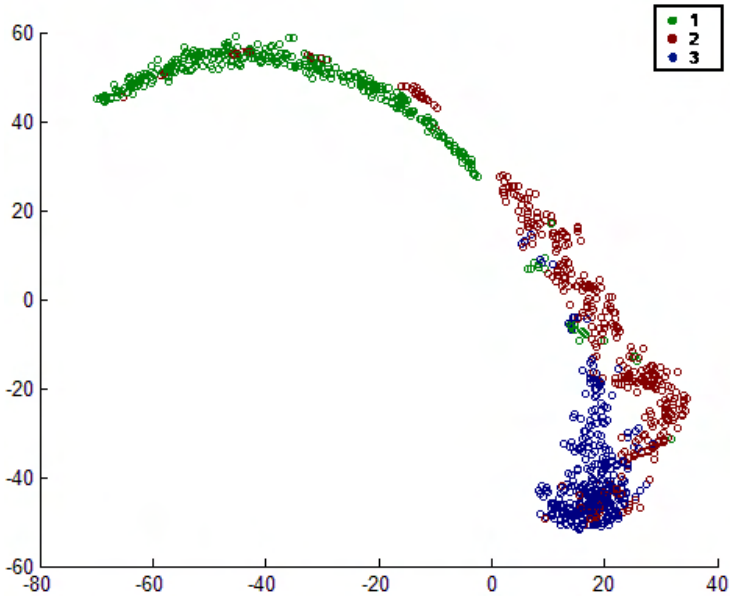

Figura 6.2: *Proyección de los vectores de características de las tres clases de emociones usando t-SNE: 1) neutral, 2) alegría and 3) tristeza*

6.3. Parámetros de la experimentación

Para los experimentos de RBM y DBN modificamos la herramienta desarrollada por Drausin Wulsin [66], lo que permitió llevar a cabo un gran número de éstos con el fin de determinar las mejores configuraciones y parámetros. Estos experimentos consistieron en diferentes combinaciones de tamaño de lote, tasa de aprendizaje, número de unidades ocultas, y número de RBMs. Cabe mencionar que la primera aproximación a esta selección general de parámetros se basó en los consejos que se encuentran en el Anexo D.

En la Tabla 6.1 se muestran los valores de los parámetros con los que se realizaron los experimentos de manera exhaustiva, con todas las posibles elecciones de los parámetros:

Parámetros	Valores
Tamaño del lote	[6, 12, 18, 24, 30, 36, 42, 48, 54, 60]
Tasa de aprendizaje	[0.01, 0.001, 0.0001, 0.00001]
Unidades ocultas	[28, 56, 84, 112, 140, 168]

Tabla 6.1: *Parámetros de configuración para el entrenamiento de las RBM y DBN con tres emociones*

Sean tl, ta y uo los vectores correspondientes a el tamaño del lote, la tasa de aprendizaje y las unidades ocultas, creamos un conjunto ordenado de vectores de experimentos P cuyos elementos son de la forma $vp = (tl_i, ta_j, uo_k)$ siguiendo el orden lexicográfico. En la tabla 6.2 se muestra un ejemplo de esta selección.

La arquitectura de la DBN para estos experimentos fue constituida por dos RBMs, para la primera se utilizaron los distintos números de unidades ocultas revisados en la Tabla 6.1 y 30 unidades visibles que son las 30 características extraídas de la señal del habla, para la segunda se utilizaron 3

	vp (1,1,1)	vp (1,1,2)	...	vp (1,2,1)	...	vp (10,4,6)
Tamaño del lote	6	6	...	6	...	60
Tasa de aprendizaje	0.01	0.01	...	0.001	...	0.00001
Unidades ocultas	28	56	...	28	...	168

Tabla 6.2: *Selección de parámetros de entrenamiento de la Tabla 6.1 en cada ejecución*

unidades ocultas y el número de unidades ocultas de la capa anterior como unidades visibles, pues como hemos explicado, la salida de la primer capa alimenta a la segunda.

La segunda capa es la que funciona como clasificador, pues cada una de las tres unidades ocultas determinará a cuál de las emociones aludidas pertenece el vector presentado. Así, la clase más probable fue considerada como la unidad con el mayor nivel de activación.

Además de estos experimentos con 30 características, otros se realizaron con los siguientes clasificadores: K-nn, árbol de decisión (DT), perceptrón multicapa (MLP) y máquinas de soporte vectorial (SVM). La implementación de los algoritmos y parámetros empleados para la configuración de los tres primeros están dados por el software Matlab [67] y no han sido modificados, para las máquinas de soporte vectorial se han utilizado los valores por defecto del software LibSVM [68]. A continuación se detallan:

- Para K-NN:
  - Se utilizó un vecino y como medida de distancia, la euclidiana
- Para DT:
  - Se utilizó el índice de la diversidad de Gini, y luego se podó a fin de obtener una mejor capacidad de generalización

- Para MLP:
  - Se entrenó con el algoritmo clásico de retro-propagación y una capa oculta con tres unidades sigmoideas
- Para SVM:
  - Se utilizó una SVM del tipo *C-SVC* con la función de kernel en base radial de grado 3 $exp(-gamma * |u - v|^2)$, gamma de $1/numCaract$, coef0 de 0, el parámetro C con valor de 1 y 0,001 como tolerancia de terminación.

También realizamos algunos experimentos 'mixtos' donde los clasificadores K-NN, DT, MLP y SVM se alimentaron con las salidas de una RBM con el objetivo de determinar si el preprocesamiento realizado por la primer RBM ayudó a mejorar el desempeño de estos otros clasificadores.

Estos experimentos con 30 características no fueron los únicos que realizamos. Utilizamos la misma metodología para experimentar con otros grupos que fueron propuestos como resultado del concurso INTERSPEECH organizado por la International Speech Communication Association, ISCA. El de los años 2009 [69] y 2010 [70] nos resultan de particular interés pues hubo una subcategoría que exploró la clasificación de emociones, el número de características para el año 2009 fue de 384 mientras que para el 2010 fue de $1,582$. Estas se pueden observar en el Anexo B sin embargo, los resultados fueron significativamente inferiores a los encontrados con nuestra propuesta.

Una vez determinados qué parámetros dieron el mejor resultado, se preparó un segundo grupo de experimentos que consistieron en ir aumentando de uno en uno el número de RBMs apiladas en la DBN hasta llegar a 15 capas con la idea de verificar si el desempeño mejoraba. Los resultados de los dos grupos de experimentos se muestran en la sección siguiente.

6.4. Resultados

Los tres mejores y tres peores resultados de los experimentos llevados a cabo con los parámetros de la Tabla 6.1 y evaluados con la partición de prueba se muestran en la Tabla 6.3 mientras que, en la Figura 6.3 se observa la comparación con todos los clasificadores, incluso aquellos que fueron alimentados con la salida de la RBM.

T. lote	T. aprendizaje	U. ocultas	% Error
54	0.0001	84	2.51
24	0.001	140	2.51
60	0.0001	56	2.51
12	0.01	56	48.39
12	0.01	168	52.69
24	0.01	28	53.76

Tabla 6.3: *Mejores y peores resultados de los experimentos presentados en la Tabla 6.1 ordenados de menor a mayor porcentaje de error*

La combinación de parámetros con los que se obtuvieron los mejores resultados fue: 84 unidades ocultas, un tamaño de partición de 54 y una tasa de aprendizaje de 0.0001. Con esta configuración en particular, la DBN alcanzó una tasa de error de 2.51 %. Cabe mencionar que estos resultados son superiores a los alcanzados por los alumnos según la prueba subjetiva que se encuentra en la Sección 5.3 de este trabajo.

Finalmente hemos utilizado esta configuración, que produjo la tasa de error de 2.51 %, con el fin de realizar un segundo conjunto de experimentos con las 15 capas apiladas de las que hemos hablado en este mismo capítulo. Los resultados que se pueden ver en la Figura 6.4 muestran que el mejor se logró con una, dos y tres capas de RBMs y luego empeoró. Una posible

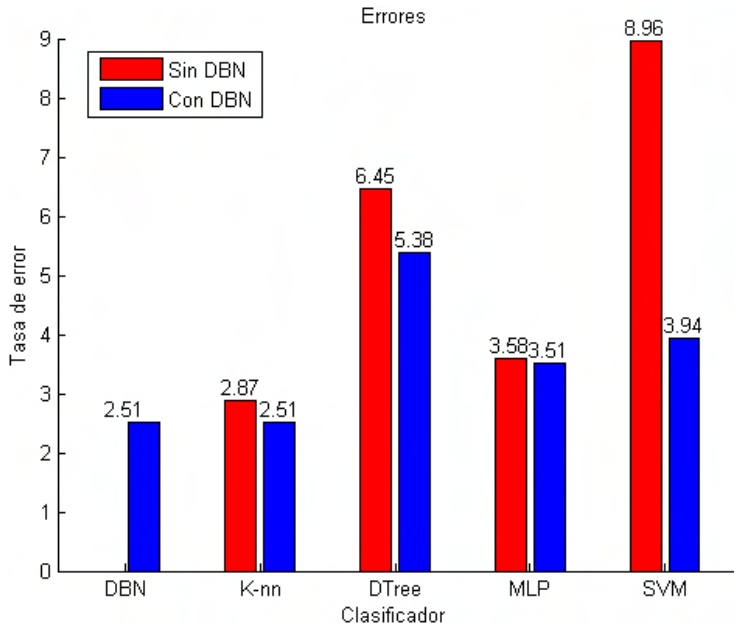

Figura 6.3: *Tasas de error de los clasificadores para tres emociones*

explicación de este resultado puede ser el uso de un pequeño subconjunto de emociones y datos, ya que el incremento en las capas de la DBN implica un aumento en el número de valores a entrenar, lo que requiere datos adicionales para estimar correctamente los parámetros, aunque es necesario ahondar en esta investigación para que podamos probar esta aseveración.

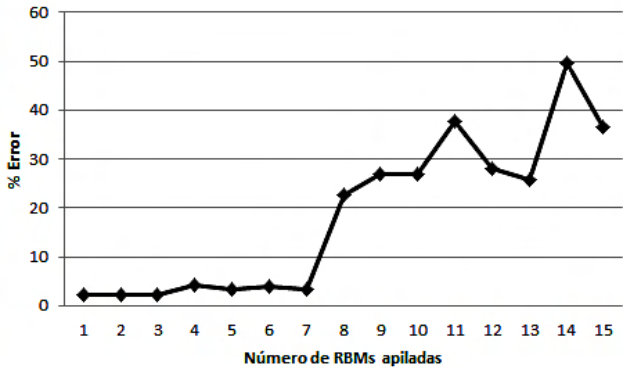

Figura 6.4: *Tasa de error contra número de RBMs apiladas*

Es importante mencionar que los errores que cometió la DBN fueron cometidos en palabras y no en oraciones completas lo que confirma los resultados empíricos presentados en el Capítulo 5.3. También vale la pena hacer notar que la combinación de la DBN y los otros clasificadores, en base a los resultados obtenidos con esta selección de parámetros, resultó en una mejora de clasificación.

En el Anexo E se encuentran las tablas con los resultados completos obtenidos al combinar los parámetros de la Tabla 6.1.

CAPÍTULO 7

CONCLUSIÓN

Las conclusiones que se exponen a continuación, permiten verificar el logro de los objetivos propuestos, y muestran que hacer uso de las máquinas restringidas de Boltzmann, es una técnica prometedora que, con base en la metodología propuesta, ha logrado obtener resultados comparativamente superiores a otros métodos de clasificación, e incluso, puede mejorar el desempeño de estos. Es con esto que el trabajo realizado en esta investigación, permite establecer las siguientes conclusiones de acuerdo a los objetivos planteados en la Sección 1.4 de este trabajo:

- En el Capítulo 2 se expuso la metodología para crear un sistema de reconocimiento de emociones a partir de la señal del habla mediante el uso de Redes de Creencia Profunda y Máquinas Restringidas de Boltzmann.
- En el Capítulo 6 se propuso hacer uso de un grupo de características que mostró ser capaz de describir correctamente las emociones en el habla (6.2).
- Los porcentajes de clasificación obtenidos en las pruebas realizadas al

sistema (Figuras 6.3 y 8.10), permiten concluir que es posible crear un sistema de reconocimiento de emociones en el habla que alcance mejores resultados que los conseguidos por los alumnos que participaron en la evaluación presentada en la Sección 5.3. De los resultados mostrados en las Secciones 8.1 y 5.4 podemos decir que aunque el trabajo con siete emociones no fue exhaustivo, el uso de las máquinas profundas de Boltzmann ofrece resultados prometedores.

- De estos resultados también se concluye que el sistema presentado en este trabajo tiene un mejor desempeño que los clasificadores tradicionales. De hecho, el preprocesamiento realizado por la primer RBM, al ser utilizado para alimentar a dichos clasificadores, parece ayudarlos a discriminar mejor entre las distintas emociones.

CAPÍTULO 8

DISCUSIÓN Y PERSPECTIVAS

En este trabajo se consideró la aplicación de las RBMs y DBNs en la tarea de reconocimiento automático de las emociones en el habla en español. Esto permitió obtener resultados comparables, y en los casos explorados, mejores que los resultados de otros clasificadores cuando los parámetros son elegidos correctamente.

A pesar de ello, debido a que el uso de las DBNs y RBMs es relativamente nuevo y más aún en el área del reconocimiento de emociones, existen pocas aproximaciones a la interpretación del funcionamiento de las capas de la red [41]. Es por esto que, como resultado de la experimentación realizada en este trabajo, en los párrafos siguientes se propone una interpretación mediante la comparación de las matrices de pesos en las diferentes capas de la DBN que fue entrenada en el Capítulo 6.

Las imágenes que se muestran a continuación son los pesos de la primer capa de la DBN, ésta fue entrenada 4 veces de la siguiente manera: con las tres emociones juntas, sólo con neutral, sólo con alegría y sólo con tristeza.

En las Figuras 8.1 a 8.4 se muestran los pesos obtenidos en dichos entrenamientos. En el eje y se tienen las 30 características extraídas, mientras que en el eje x están las 84 unidades ocultas.

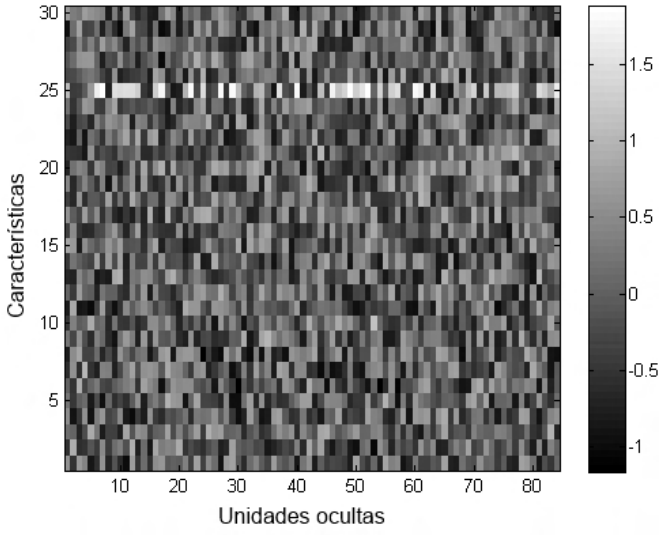

Figura 8.1: *Pesos de la primera RBM después de ser entrenada con las 3 emociones; neutra, alegría y tristeza*

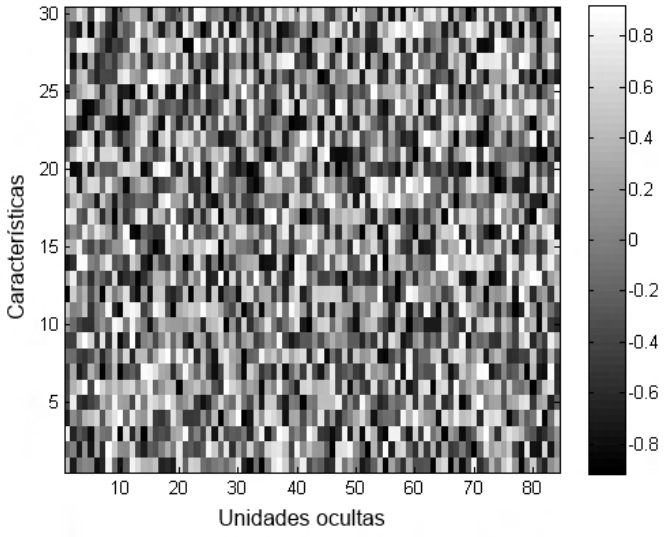

Figura 8.2: *Pesos de la primera RBM después de ser entrenada con neutral*

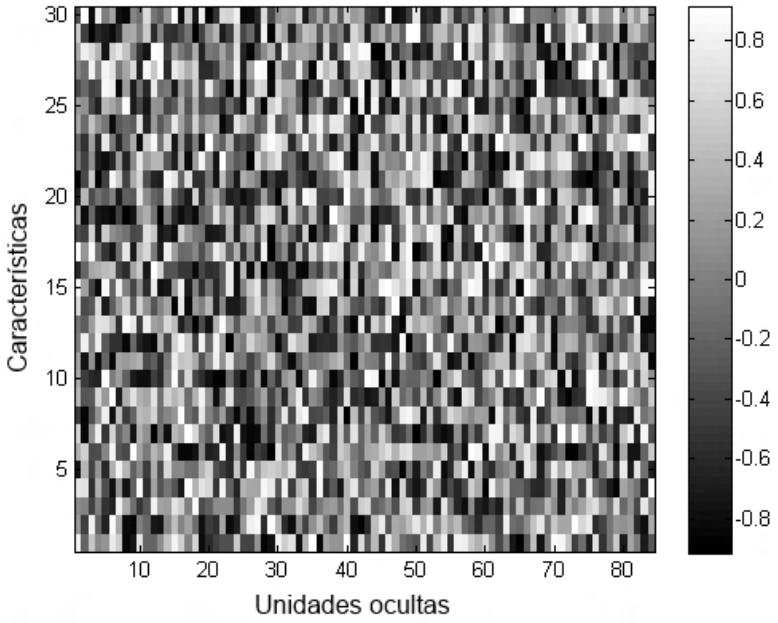

Figura 8.3: *Pesos de la primera RBM después de ser entrenada con alegría*

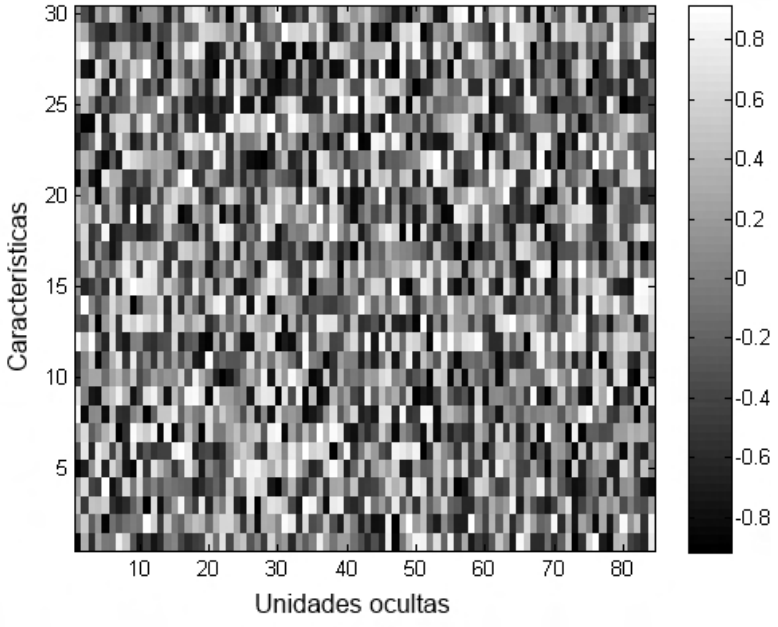

Figura 8.4: *Pesos de la primera RBM después de ser entrenada con tristeza*

De estas imágenes podemos obtener algunas interpretaciones, por ejemplo las intensidades que vemos en las imágenes van gradualmente aumentando de neutral a alegría. Un parámetro que ayuda a determinar ésto es la media absoluta de los pesos como se puede apreciar en la Tabla 8.1 además, en la Figura 8.1 podemos observar que F_0, que es la característica 25, posee un color predominantemente claro en comparación con las otras características. Esto puede significar que es un elemento valioso en la elección de parámetros aunque no se ha evidenciado en la experimentación de manera importante.

Emoción de entrenamiento	Media absoluta
Neutra, alegría y tristeza	0.47
Neutra	0.40
Alegría	0.49
Tristeza	0.45

Tabla 8.1: *Medias absolutas de los pesos en la primer capa de las RBMs*

Esta interpretación de la primer capa de la DBN deja ver la importancia de continuar investigando el funcionamiento de las RBMs, es primordial extender esta experimentación para abarcar las siete emociones disponibles en la base de datos así como la incorporación de distintos hablantes a las pruebas. Por este motivo en la Sección siguiente (8.1), se da una primera aproximación a más emociones.

Las imágenes que aparecen a continuación siguen el mismo orden que las anteriores pero no muestran los pesos, sino los valores de activación de cada unidad oculta de la primer RBM, es necesario realizar más pruebas para poder proponer una interpretación, pues no parece haber un patrón identificable para cada una de las emociones.

Aquí se muestra en el eje x las 84 unidades ocultas y en el eje y los vectores de entrenamiento.

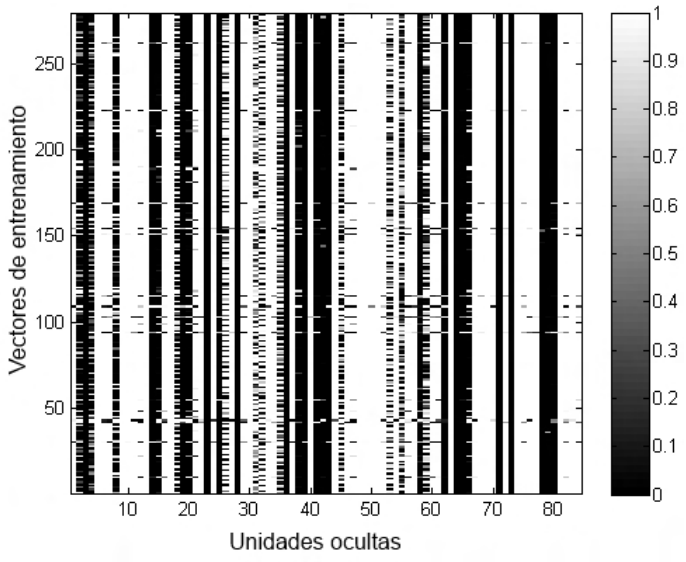

Figura 8.5: *Valores de activación de la primer RBM presentando tres emociones; neutra, alegría y tristeza*

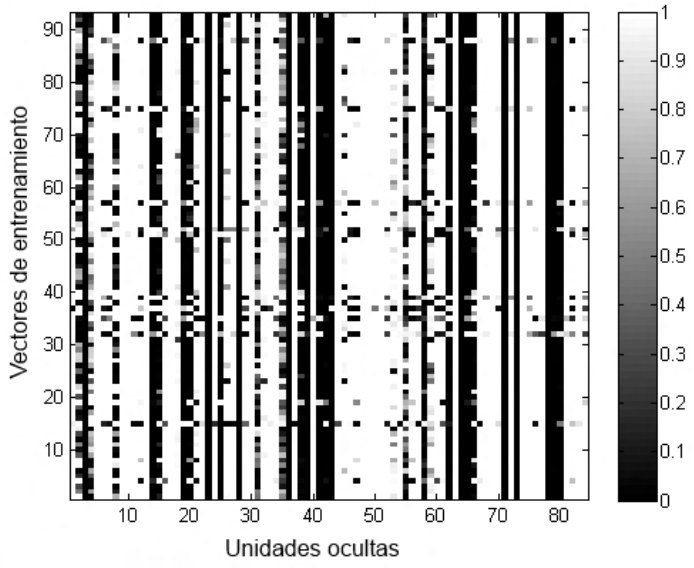

Figura 8.6: *Valores de activación de la primer RBM para la emoción: neutra*

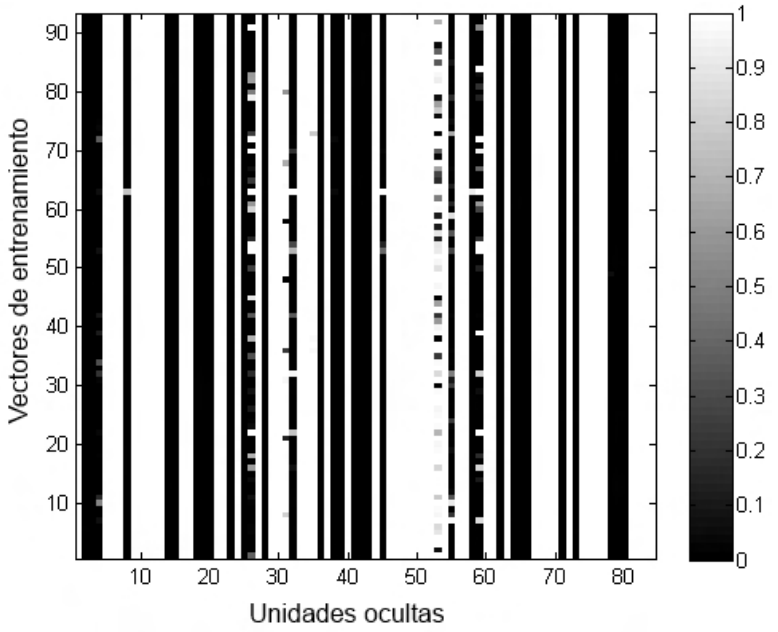

Figura 8.7: *Valores de activación de la primer RBM para la emoción: alegría*

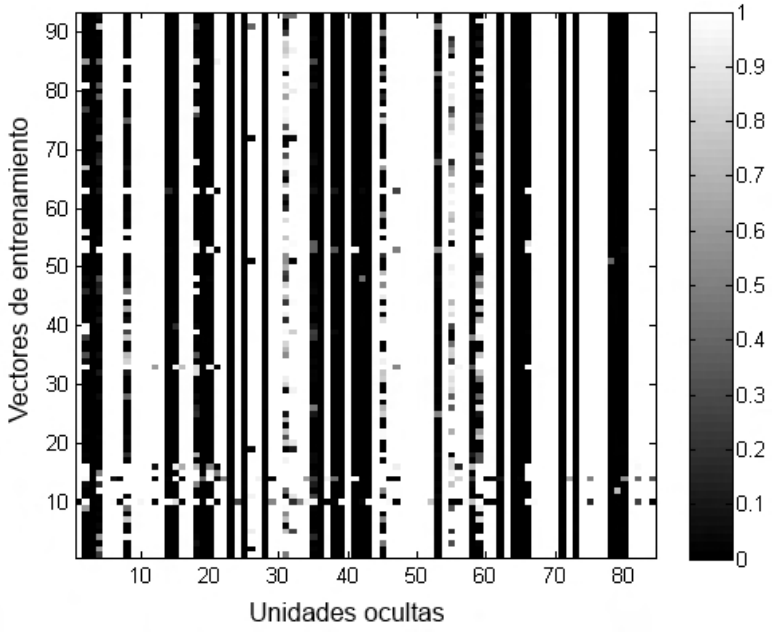

Figura 8.8: *Valores de activación de la primer RBM para la emoción: tristeza*

Actualmente estamos trabajando en la posibilidad de que si presentamos un patrón, por ejemplo de tristeza, a una red entrenada únicamente con alegría, podremos medir mediante algún tipo de distancia, la cercanía existente entre las activaciones de las unidades ocultas, para poder generar un cluster alrededor de estas imágenes. Esta idea no es desconocida, se han utilizado distancias o medidas de error en los Códigos de Respuesta Rápida (QR code) [71], en la Figura 8.9 se da un ejemplo de estos códigos.

Esta idea puede ser 'parametrizada' con distintas medidas de distancias y otros valores además de las activaciones de las unidades ocultas, por ejemplo directamente de los pixeles de la imagen.

Figura 8.9: *Código de Respuesta Rápida (QR) de la dirección móvil de Wikipedia; http://en.m.wikipedia.org*

8.1. Extensión a siete emociones

Aún resta mucho por hacer para poder decir algo concluyente sobre la interpretación de las capas que fueron entrenadas utilizando tres emociones, sin embargo, gracias a la experiencia adquirida a lo largo de la experimentación y a las heurísticas disponibles, hemos podido obtener algunos primeros resultados favorables al experimentar con las siete emociones.

Estas tasas de error son, cuando menos, comparables con los resultados vistos en el Anexo A y en la Sección 5.4. Creemos que tras realizar ajustes a los parámetros se podrán disminuir, pero como dijimos, aún resta mucho trabajo por hacer. Para obtener estos resultados se siguió el mismo procedimiento que vimos en el Capítulo 6 con los siguientes parámetros para el entrenamiento de las RBMs y DBN:

Parámetros	Valores
Tamaño de la partición	56
Tasa de aprendizaje	0.001
Unidades ocultas	112
Número de RBMs	3

Tabla 8.2: *Parámetros de configuración para el entrenamiento de las RBMs y DBNs para siete emociones*

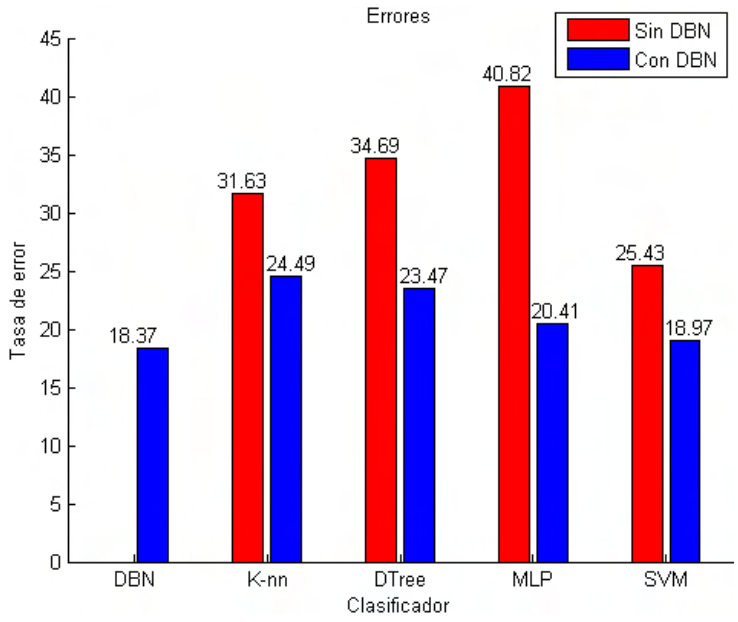

Figura 8.10: *Tasas de error de los clasificadores para siete emociones*

También se ha considerado realizar la experimentación necesaria para determinar la influencia que podría tener el aplicar estas técnicas a otros idiomas o entonaciones, como es el caso del español de México cuya base de datos está siendo creada por el grupo al que pertenezco. No obstante la base de datos para español de México no ha sido terminada, en la siguiente Sección se presenta la extensión al idioma Alemán.

8.2. Extensión a otros idiomas

Para realizar los experimentos con el idioma Alemán se ha utilizado la *Berlin Database of Emotional Speech* [1], que es una base de datos creada en el departamento de Acústica Técnica de la Universidad Técnica de Berlín, con el propósito de estudiar el habla con emociones y la síntesis de voz. Las frases expresan seis emociones más neutral hablas por actores, cinco hombres y cinco mujeres que enunciaron diez frases.

8.2.1. Estructura

Las emociones que se muestran a continuación son prácticamente las mismas que las grabadas en la base de datos en español descrita en el Capítulo 5. Estas variaciones en el estilo son: neutral, alegría, tristeza, enojo, miedo, disgusto y aburrimiento como se muestra en la Tabla 8.3.

Alemán	
A = enojo	F = miedo
B = aburrimiento	H = alegría
D = disgusto	S = tristeza
N = neutral	

Tabla 8.3: *Lista de emociones en Alemán*

8.2.2. Corpus

El corpus está compuesto por las 10 frases que se muestran en la Tabla 8.4. Como se mencionó anteriormente, estas frases fueron expresadas por 10 actores en siete emociones, con esto tenemos alrededor de $7 * 10 * 10 = 700$ archivos de audio, de los cuales sólo 500 están disponibles para descarga.

Frases en Alemán	Traducción
Der Lappen liegt auf dem Eisschrank	El mantel está sobre en la nevera
Das will sie am Mittwoch abgeben	Ella lo entregará el miércoles
Heute abend könnte ich es ihm sagen	Esta noche yo podría decirle
Das schwarze Stück Papier befindet sich da oben neben dem Holzstück	La hoja de papel negro se encuentra allí, a un lado del trozo de madera
In sieben Stunden wird es soweit sein	En siete horas lo será
Was sind denn das für Tüten, die da unter dem Tisch stehen?	¿Qué pasa con las bolsas que estan allí debajo de la mesa?
Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter	Ellos simplemente lo llevaron arriba y ahora van a bajarlo de nuevo
An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht	Los fines de semana siempre voy a casa y veo a Agnes
Ich will das eben wegbringen und dann mit Karl was trinken gehen	Me limitaré a desechar este y luego iré a tomar algo con Karl
Die wird auf dem Platz sein, wo wir sie immer hinlegen	Será en el lugar donde siempre la guardamos

Tabla 8.4: *Lista de frases y su traducción al español*

8.2.3. Evaluación subjetiva

Los creadores de esta base ponen a nuestra disposición el estudio que realizaron para evaluarla y que se muestra a continuación [1]. Dicha evaluación consistió en realizar pruebas subjetivas donde participaron 20 oyentes. A estos sujetos se les presentaron los audios de manera aleatoria con sólo una oportunidad para escucharlos antes de clasificarlos.

En la Figura 8.11 se muestra la tasa media de reconocimiento. Las lineas que conectan las barras indican una diferencia significativa entre emociones.

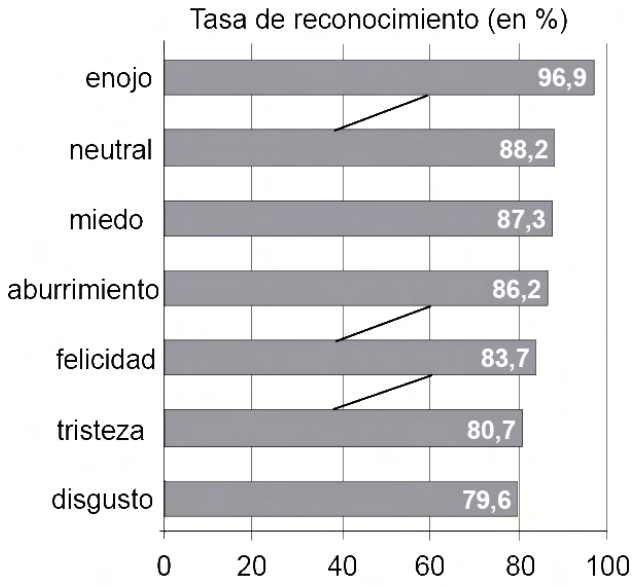

Figura 8.11: *Tasa de reconocimiento para la base de datos de Berlín, tomada de [1]*

8.2.4. Evaluación mediante DBN

El objetivo de la experimentación descrita en esta sección es determinar si las características elegidas para el español (6.2) son útiles para clasificar las emociones en otro idioma, en particular para el Alemán. Por este motivo no se discuten otras características ni la experimentación exhaustiva que se realizó con el Español.

El primer experimento realizado consistió en presentar todos los vectores de características de tres emociones: neutral, alegría y tristeza, las mismas que para español, a la DBN entrenada previamente en la Sección 6.4. Los resultados se muestran a continuación.

	N	H	S	
N	32	26	21	79
H	0	70	9	79
S	65	4	10	79
	97	100	40	237

Tabla 8.5: *Matriz de confusión con la red pre-entrenada. Los valores en las columnas representan el número de oraciones reconocidas contra las emociones reales en cada fila.*

Como la red estaba previamente entrenada para reconocer los patrosnes del habla en español, es de esperarse que la tasa de error fuera tan alta como el 52,74 % que se obtuvo. Con esto queda claro que el idioma es de vital importancia en el entrenamiento de una DBN pues las características fonético-acústicas de los distintos idiomas es variada.

El segundo experimento tuvo como objetivo verificar si las características seleccionadas permiten discriminar las emociones aún cuando el idioma es diferente. Para lograrlo se realizó el entrenamiento de la red haciendo uso de

los 237 vectores de características separados en 156 (70 %) entrenamiento, 63 (25 %) prueba y 18 (5 %) validación. El grupo de parámetros es el mismo que el utilizado en la Sección 6.4.

Los resultados expuestos en la Tabla 8.6 en comparación con los de la tabla anterior, dejan ver que las características seleccionadas son suficientemente buenas como para discriminar entre emociones sin importar si son en Español o en Alemán. Existen algunas salvedades que debemos hacer notar para interpretar estos resultados, la primera de ellas es el número reducido de vectores en la partición de entrenamiento y la segunda es que no se ha realizado una exploración exhaustiva para determinar los parámetros de entrenamiento para las RBMs.

	N	H	S	
N	19	2	0	21
H	1	19	1	21
S	0	0	21	21
	20	21	22	63

Tabla 8.6: *Matriz de confusión tras el entrenamiento de la DBN. Los valores en las columnas representan el número de oraciones reconocidas contra las emociones reales en cada fila.*

Creemos que con un ajuste fino de los parámetros se puede mejorar la tasa de error de 6,35 % aunque como mencionamos, el objetivo era determinar la pertinencia de las características y no los parámetros por si mismos.

BIBLIOGRAFÍA

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," *INTERSPEECH*, pp. 1517–1520, 2005.

[2] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput*, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[4] M. Schubiger, *English Intonation. Its Form and Function ISBN-13: 978-3484400184*. Niemeyer Verlag, 1958.

[5] J. O'Connor and G. F. Arnold, *Intonation of Colloquial English ISBN-13: 978-0582523890*. Prentice Hall Press, 1973.

[6] M. Benzeghiba, D. Renato, D. Olivier, D. Stephane, E. Theodora, J. Denis, and F. Luciano, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[7] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, pp. 53–56, 1968.

[8] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[9] S. Marc and C. R. et.al, "Feeltrace: An instrument for recording perceived emotion in real time," *ISCA Workshop on Speech and Emotion*, pp. 19–24, 2000.

[10] H. Lovheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Medical Hypotheses*, vol. 78, no. 2, pp. 341–349, 2012.

[11] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process*, no. 18, pp. 32–80, 2001.

[13] E. Paul, V. Wallace, and C. Joseph, "Facial action coding system: The manual on cd rom. a human face," 2002.

[14] Tsoumakas, K. Grigorios, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*, p. 325, 2008.

[15] Zentner, Marcel, D. Grandjean, and K. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[16] S. Girard, M. E. C. Echeagaray, J. G. Sanchez, Y. Hidalgo-Pontet, L. Zhang, W. Burleson, and K. VanLehn, "Defining the behavior of an affective learning companion in the affective meta-tutor project," *Artificial Intelligence in Education (AIED)*, pp. 21–30, 2013.

[17] R. W. Picard and J. Healey, "Affective wearables," *Personal Technologies*, vol. 1, pp. 231–240, 1997.

[18] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist, "Integrating models of personality and emotions into lifelike characters," *Affective interactions*, vol. 1814, pp. 150–165, 2000.

[19] M. vant Wout, R. S. Kahn, A. G. Sanfey, and A. Aleman, "Affective state and decision-making in the ultimatum game," *Experimental Brain Research*, vol. 169, pp. 564–568, 2006.

[20] C. Magerkurth, A. D. Cheok, R. L. Mandryk, and T. Nilsen, "Pervasive games: bringing computer entertainment back to the real world," *Computers in Entertainment (CIE)*, vol. 3, p. 4A, 2005.

[21] A. Hassan, *On Automatic Emotion Classification Using Acoustic Features*. University of Southampton: Electronics and Computer Science, 2012.

[22] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: design and collection of a multilingual emotional speech database," *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC*, pp. 2024–2028, 2002.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[24] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech communication*, vol. 40, no. 1, pp. 33–60, 2003.

[25] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, pp. 1162–1181, 2006.

[26] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, no. 44, pp. 572–587, 2011.

[27] F. J. C. Serena and F. J. Cantero, *Teoría y análisis de la entonación*, vol. 54. Edicions Universitat Barcelona, 2002.

[28] J. Krajewski and B. Kroger, "Using prosodic and spectral characteristics for sleepiness detection," *Interspeech Proceedings*, vol. 8, pp. 1841–1844, 2007.

[29] G. Stemmer, *Modeling variability in speech recognition*. Logos-Verlag, 2005.

[30] C. Busso, L. Sungbok, and N. Shrikanth, "Using neutral speech models for emotional speech analysis," *Interspeech*, pp. 2225–2228, 2007.

[31] J. Laver, "The phonetic description of voice quality," *Cambridge University Press*, 1980.

[32] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1, pp. 189–212, 2003.

[33] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Searching

for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language*, no. 25, pp. 4–28, 2011.

[34] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," *Language Technologies*, pp. 240–245, 2006.

[35] M. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," *Interspeech*, pp. 2973–2976, 2007.

[36] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *ICASSP*, pp. 577–580, 2004.

[37] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," *ICSLP*, 2004.

[38] J. Ang, R. Dhillon, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," *Interspeech*, pp. 2037–2040, 2002.

[39] H. Hu, M.-X. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," *ICASSP*, vol. 4, p. 413, 2007.

[40] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," *ICASSP, IEEE*, pp. 5688–5691, 2011.

[41] E. Schmidt and Y. Kim, "Learning emotion-based acoustic features with deep belief networks," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 65–68, 2011.

[42] C. Williams and K. Stevens, "Vocal correlates of emotional states, speech evaluation in psychiatry," *Grune and Stratton*, pp. 189–220, 1981.

[43] H. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust. Speech Signal Process*, vol. 5, no. 28, pp. 599–601, 1990.

[44] G. Izzo, "Multiresolution techniques and emotional speech," *PHYSTA Project Report*, 1998.

[45] F. J. S. Campos, *Modelos Ocultos de Markov: Del Reconocimiento de Voz a la Musica, 978-1847536778.* LuluPress, 2009.

[46] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[47] E. Kandel, J. Schwartz, and T. Jessell, "Principles of neural science," *McGraw-Hill Medical*, 2000.

[48] H. Ghashghaei, C. Hilgetag, and H. Barbas, "Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala," *NeuroImage*, vol. 34, no. 3, pp. 905–923, 2007.

[49] M. Bar, R. Tootell, D. Schacter, D. Greve, B. Fischl, J. Mendola, B. Rosen, and A. Dale, "Cortical mechanisms speci," *Neuron*, vol. 29, no. 2, pp. 529–535, 2001.

[50] Ackley, Hinton, and Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive Science*, vol. 9, no. 1, pp. 147–169, 1985.

[51] Hinton and Sejnowski, "Optimal perceptual inference," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 448–453, 1983.

[52] G. Hinton, "A practical guide to training restricted boltzmann machines," *Department of Computer Science, University of Toronto*, 2010.

[53] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-Moisy, "An introduction to deep learning," *inproceedings*, 2011.

[54] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, no. 50, pp. 5–43, 2003.

[55] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[56] E. Chen, "Edwin chen's blog." `http://blog.echen.me/2011/07/18/introduction-to-restricted-boltzmann-machines/`.

[57] G. Hinton, "Learning multiple layers of representation," *TRENDS in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[58] A. Batliner, S. Steidl, B. Schuller, and D. Seppi, "The hinterland of emotions: Facing the open-microphone challenge," *ACII*, 2009.

[59] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," *INTERSPEECH*, pp. 2679–2682, 2001.

[60] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," *Prentice Hall PTR*, 1993.

[61] E. Albornoz, D. Milone, and H. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech and Language*, vol. 25, pp. 556–570, 2011.

[62] F. Eyben, M. Wollmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," *ACM Multimedia (MM)*, pp. 1459–1462, 2010.

[63] L. der Maaten, E. Postma, and H. den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, pp. 1–41, 2009.

[64] I. Jolliffe, *Principal component analysis*. Springer verlag, 2002.

[65] L. der Maaten and G. Hinton, "Visualizing high- dimensional data using t-sne," *Journal of Machine Learning Research*, pp. 2579–2605, 2008.

[66] D. Wulsin, "Dbn toolbox v1.0, department of bioengineering, university of pennsylvania," 2010.

[67] Mathworks, "Matlab users guide (r2013a)." `http://www.mathworks.com`, 2013.

[68] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, April 2011.

[69] ISCA, "The interspeech 2009 emotion challenge," *ISCA*, pp. 312–315, 2009.

[70] ISCA, "The interspeech 2010 paralinguistic challenge," *ISCA*, 2010.

[71] ISO-IEC-18004-2006, "Information technology. automatic identification and data capture techniques. qr code 2005 bar code symbology specification." ISBN:978-0-580-67368-9, 2007.

[72] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech : A review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[73] M. B.S., S. Philippe, and S. Thomas, *Introduction to MPEG-7: Multimedia Content Description Interface, ISBN 0-471-48678-7*. Wiley and Sons, 2002.

[74] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson Education, 2011.

Apéndices

APÉNDICE A

OTROS TRABAJOS SOBRE EL RECONOCIMIENTO DE EMOCIONES

En la Figura 2.1 se puede ver que el primer componente en un sistema de reconocimiento de emociones en el habla es la disponibilidad de las bases de datos emocionales. Distintos autores han enumerado varias bases de datos de voz que han sido utilizadas por muchos investigadores que participan en el reconocimiento de emociones [24, 25, 26]. Tres diferentes tipos de bases de datos se pueden observar en la literatura: las que contienen discurso simulado o actuado en la que las emociones se expresan deliberadamente por profesionales, otras, de habla natural o espontánea en la que todas las grabaciones se realizan en el entorno del mundo real y, también están las de habla inducida donde se controló el ambiente de forma tal que los participantes expresaran, mediante la provocación, sus emociones.

Es importante mencionar que aunque se cuenta con numerosos estudios de clasificación de emociones en la voz, existen muy pocos que utilizan las arquitecturas profundas y menos aún los que las combinan con RBMs como lo hicieron André Stuhlsatz et al., en el 2011 [40] que se encuentra en la posición

9 de esta tabla que muestra los detalles de las bases de datos utilizadas en la actualidad. Una revisión exhaustiva se puede encontrar en [26, 33, 40, 72, 41, 21].

Características	Bases de datos	Modelos y clasificadores	Emociones	% Desempeño
Reconocimiento de emociones utilizando características espectrales				
MFCC y LFPC	Películas y clips de audio de programas de televisión	Gaussian Mixture Model-GMM, Redes Neuronales-NN, NN Difusas	Enojo, miedo, disgusto, alegría, tristeza y neutral	Por encima del humano
MFCC	LDC-Inglés, Emo-DB	Support Vector Machine-SVM	Enojo, miedo, disgusto, alegría, tristeza y neutral para LDC y Emo-DB	44.5% y 78.2% para cada base de datos respectivamente
MFCC, Spectral contrast, Statistical spectrum descriptors, Autocorrelation of chroma, Echo Nest Timbre (ENT, propietario)	Moodswings Lite	DBN y Multiple Linear Regression-MLR	Videoclips de canciones de 15 segundos	No reportado
Reconocimiento de emociones utilizando características prosódicas				
Picos y depresiones en el pitch y perfiles de energía, duración de las pausas y ráfagas	Los autores hicieron su propia base de datos	Análisis Discriminante Lineal	Enojo, miedo, alegría y tristeza	55%
Características prosódicas basadas en el suavizado de los contornos del pitch	Los autores hicieron su propia base de datos en inglés	Voto de especialistas	Enojo, miedo, felicidad y tristeza	80%
Reconocimiento de emociones utilizando características espectrales y prosódicas				
87 características relacionadas al pitch y parámetros espectrales	Base de datos de emociones Danesa	K-medias y clasificadores de Bayes	Enojo, felicidad, neutral, tristeza y sorpresa	51.6%
Prosodia: patrones de entonación y poder Espectral: LPCs y sus coeficientes delta	Basada en actores, 100 hablantes Japonenses	Redes Neuronales	Enojo, disgusto, neutral, tristeza, sorpresa y burla	50%
Prosodia: pitch, energía, delta de las energías mel y características del pitch Espectral: formantes, MFCCs y sus características delta	SUSAS, AIBO	Análisis discriminante cuadrático	Estresada y neutral para SUSAS y cuatro clasificaciones para AIBO	96.3%, 70.1% y 42.3% para cada base de datos respectivamente
39 características funcionales y 56 LLDs	ABC, AVIC, DES, EMOD, eNTER, SAL, Smart, SUSAS, VAM. (inglés, alemán y danés)	Mahalanobis, SVM Polinomial, GerDA*	No estandarizadas entre las bases de datos, número de expresiones; 430, 3002, 419, 494, 1277, 1692, 3823, 3593, 946	80.6%, 85.5%, 90.3%, 97.6%, 80.8%, 66.4%, 89.4%, 83.3%, 92.3%, 82.3% para cada base de datos respectivamente
Duración, Energía, Tono, Espectro, Cepstrum, Calidad de la voz, Wavelets, Bag of words, Part-of-speech, Higher semantics,	AIBO	NN, SVM, Random Forests-RF, Regresión Lineal, Discriminante Lineal, Bayes ingenuo, Clasificadores basados en reglas	Alegre, Sorprendido, Maternal, Neutral, Descanso, Aburrido, Enfático, Impotente, Susceptible (irritado), Reprimenda, Enojado	55.3%
Tono, Energía, Formantes, Calidad de la voz, LPC, MFCC, LFPC	LDC, Berlin DB, Danish DB, Natural, ESMBS, INTERFACE, KISMET, BabyEars, SUSAS, MPEG-4, Beihang Univ., FERMUS III, KES, CLDC, Hao Hu et al., Amir et al., Pereira	Hidden Markov Model-HMM, GMM, NN, SVM	Distintas	75.5-78.5%, 74.83-81.94%, 51.19-52.82%, 75.45-81.29% para cada base de datos respectivamente
Pitch, Energía, ZCR, MFCC, Formantes, Epectralesy Calidad de la voz	DES, Berlin, Serbian, Aibo-Mont, Aibo-ohm	SVM lineal	Neutral, Enojo, Felicidad. Tristeza, sorpresa, Miedo, Aburrimiento, Disgusto	69.6%, 89.2%, 94.2%, 49.9%, 53.3% para cada base de datos respectivamente
Pitch, Energía, ZCR, MFCC, Formantes, Epectralesy Calidad de la voz	DES, Berlin, Serbian, Aibo-Mont, Aibo-ohm	Jerárquico	Neutral, Enojo, Felicidad. Tristeza, Sorpresa, Miedo, Aburrimiento, Disgusto	70%, 90.6%, 94.7%, 65.8%, 61.9% para cada base de datos respectivamente
Fuente de excitación: simetría glotál Espectral: características de los MFCCs	Base de datos recolectada de 10 hablantes	Bosque de camino óptimo	Enojo, felicidad, neutral y tristeza	97%
Reconocimiento de emociones utilizando características acústicas, prosódicas y lingüísticas				
Información acústica, prosódica y lingüística	Base de datos en inglés y alemán utilizando 10 hablantes y 2829 expresiones	GMM, SVM, NN y redes de creencias	Enojo, disgusto, miedo, alegría, neutral, y sorpresa	92%
Información acústica, idioma y discurso	Base de datos de una central telefónica con 1187 llamadas	Discriminante lineal y K-medias	Emociones negativas y no negativas	82% (hombre), 88% (mujer), 82% (hombre) y 79% (mujer) para cada base de datos respectivamente

APÉNDICE B

GRUPOS DE CARACTERÍSTICAS ALTERNATIVOS

Desde 1998 la International Speech Communication Association (ISCA), ha organizado el concurso INTERSPEECH, los realizados en los años 2009 [69] y 2010 [70] nos son de particular interés pues existió una subcategoría que exploró la clasificación de emociones.

En esta subcategoría de clasificación, los participantes pudieron utilizar un amplio conjunto de características acústicas que fueron proporcionados por los organizadores. Estos grupos son los que se describen a continuación, las del 2009 en el Grupo 1 y las del 2010 en el Grupo 2.

Los Descriptores de Bajo Nivel (LLD), son un conjunto de características sonoras establecidas en el estándar MPEG-7 [73], éstos miden varias características del sonido que sirven como una representación compacta del audio analizado. Si se desean conocer las características que definen todo el segmento hablado, a este conjunto de LLDs se les pueden aplicar ciertas funciones estadísticas que llamamos funcionales.

B.1. Grupo 1

En este grupo encontramos 384 características; (16 LLDs + 16 delta) * 12 funcionales:

Grupo 1

16 descriptores de bajo nivel	
pcm RMS energy	Root mean square signal frame energy
MFCC	1-12 coeficientes cepstrales en la escala de Mel
pcm_zcr	Cruces por cero de la señal, basado en ventanas
voiceProb	La probabilidad de sonoridad calculada a partir de la ACF
F_0	La frecuencia fundamental calcula a partir del cepstrum
12 funcionales	
Max	El valor máximo del contorno
Min	El valor mínimo del contorno
Rango	(Máximo-Mínimo)
Maxpos	La posición absoluta del valor máximo, basado en ventanas
Minpos	La posición absoluta del valor mínimo, basado en ventanas
Amean	La media aritmética del contorno
linregc1	La pendiente (m) de una aproximación lineal del contorno
linregc2	El desplazamiento (t) de una aproximación lineal del contorno
linregerrQ	El error cuadrático calculado como la diferencia de la aproximación lineal y el contorno real
stdDev	La desviación estándar de los valores en el contorno
Skewness	La asimetría, momento de 3er orden
Kurtosis	La curtosis, momento de 4° orden

B.2. Grupo 2

En este grupo encontramos $1,582$ características; $[(34\ LLDs{+}34\ deltas)*21\ funcionales + 19\ funcionales*(4\ pitch-LLD+4\ deltas)]+2$ características.

Grupo 2

34 descriptores de bajo nivel	
pcm loudness	La sonoridad como la intensidad normalizada elevada a una potencia de 0,3
MFCC	0-14 coeficientes cepstrales en la escala de Mel
logMel-Freq-Band	Potencia logarítmica de las bandas de frecuencia-Mel 0-7 distribuidas en un rango de 0 a 8 kHz
lspFreq	Las 8 frecuencias de líneas espectrales calculadas a partir de 8 coeficientes LPC.
F0finEnv	La evolvente del contorno suavizado de la frecuencia fundamental.
voicing-FinalUn-clipped	La probabilidad de sonoridad de la frecuencia fundamental final. Unclipped significa que no se pone a cero cuando se cae por debajo del umbral de sonoridad.
21 funcionales	
Maxpos	La posición absoluta del valor máximo, basado en ventanas
Minpos	La posición absoluta del valor mínimo, basado en ventanas
Amean	La media aritmética del contorno
linregc1	La pendiente (m) de una aproximación lineal del contorno
linregc2	El desplazamiento (t) de una aproximación lineal del contorno
linrege-rrA	El error lineal calculado como la diferencia de la aproximación lineal y el contorno real

21 funcionales, continuación	
linregerrQ	El error cuadrático calculado como la diferencia de la aproximación lineal y el contorno real
stdDev	La desviación estándar de los valores en el contorno
Skewness	La asimetría, momento de 3er orden
Kurtosis	La curtosis, momento de 4° orden
quartile1	El primer cuartil, 25 % percentil
quartile2	El primer cuartil, 50 % percentil
quartile3	El primer cuartil, 75 % percentil
iqr1-2	El rango intercuartil: cuartil2 - cuartil1
iqr2-3	El rango intercuartil: cuartil3 - cuartil2
iqr1-3	El rango intercuartil: cuartil3 - cuartil1
percentile1.0	El mínimo valor del contorno, representado por el percentil 1 %.
percentile99	El máximo valor del contorno, representado por el percentil 99 %.
pctlrange0-1	El rango de valores de la señal 'max-min' representado por el rango 1 % - 99 % percentil.
uplevelti-me75	El porcentaje de tiempo que la señal está por encima (75 % * rango + min).
uplevelti-me90	El porcentaje de tiempo que la señal está por encima (90 % * rango + min).
4 pitch-LLD	
F0final	El suavizado del contorno de frecuencia fundamental
jitterLocal	La variación local o desviación en el período, ventana a ventana
jitterDDP	El diferencial de variación de ventana a ventana
shimmerLocal	La desviación de amplitud entre los períodos de pitch, ventana a ventana

19 funcionales aplicados a '4 pitch-LLD'	
Maxpos	La posición absoluta del valor máximo, por ventana
Minpos	La posición absoluta del valor mínimo, por ventana
Amean	La media aritmética del contorno
linregc1	La pendiente (m) de una aproximación lineal del contorno
linregc2	El desplazamiento (t) de una aproximación lineal del contorno
linregerrA	El error lineal calculado como la diferencia de la aproximación lineal y el contorno real
linregerrQ	El error cuadrático calculado como la diferencia de la aproximación lineal y el contorno real
stdDev	La desviación estándar de los valores en el contorno
Skewness	La asimetría, momento de 3er orden
Kurtosis	La curtosis, momento de 4° orden
quartile1	El primer cuartil, 25 % percentil
quartile2	El primer cuartil, 50 % percentil
quartile3	El primer cuartil, 75 % percentil
iqr1-2	El rango intercuartil: cuartil2 - cuartil1
iqr2-3	El rango intercuartil: cuartil3 - cuartil2
iqr1-3	El rango intercuartil: cuartil3 - cuartil1
percentile 99	El máximo valor del contorno, representado por el percentil 99 %.
up level time 75	El porcentaje de tiempo que la señal está por encima (75 % * rango + min).
up level time 90	El porcentaje de tiempo que la señal está por encima (90 % * rango + min).

APÉNDICE C

MÉTODOS DE EXTRACCIÓN DE CARACTERÍSTICAS

Para estimar estas características, se toma como base la ventana:

$$f_s(n;m) = s(n)w(m-n) \tag{C.1}$$

Donde $s(n)$ es la señal de voz y $w(m-n)$ es la ventana de longitud N_w terminando en la muestra m.

C.1. Frecuencia fundamental

Para calcular la frecuencia fundamental se analizarán tres métodos:

- El método del espectro
- El método de auto-correlación
- El método del cepstrum

C.1.1. Método del espectro

Mirando el espectro, la frecuencia fundamental será generalmente el pico más a la izquierda. Sin embargo, para calificar como fundamental, este tono debe tener una relación armónica específica a los otros componentes de la señal muestreada.

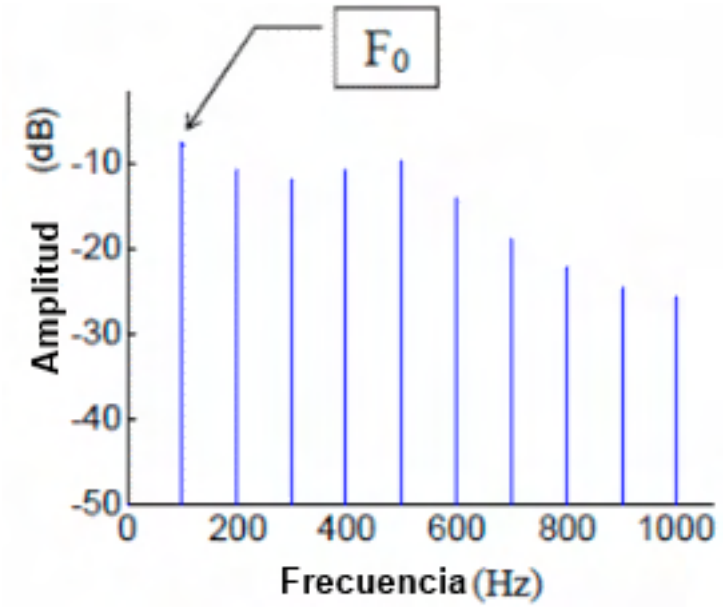

Figura C.1: *F_0, Amplutud contra Hz*

La relación es que cada tono en la señal debe ser un múltiplo de la frecuencia fundamental. Por lo tanto, si se encuentran tres picos, uno a 100 Hz, uno a 200 Hz y otro a 300 Hz, el tono a 100 Hz es el fundamental, donde el segundo armónico se encuentra a 200 Hz, el tercer armónico a 300 Hz y el quinto armónico a 500 Hz.

Sin embargo, encontrar los picos en el espectro no siempre resulta sencillo. Otra aproximación es la de tomar una porción periódica de la señal y medir

el periodo T_0. Teniendo en cuenta que la frecuencia fundamental se encuentra en $F_0 = \frac{1}{T_0}$.

Es por esto que para encontrar F_0 es necesario que la señal sea periódica, esto quiere decir que se repite en el tiempo. Así pues, un periodo es la unidad de repetición mínima en la señal. Esto quiere decir que para los múltiplos de un periodo T el valor de la señal es el mismo por lo que $F_0 = \frac{1}{T_0}$ se cumple cuando $x(t) = x(t+T) = x(t+2T) = x(t+3T) = \ldots$ donde $x(t)$ es la onda.

Por ejemplo, el sonido de la letra A en la Figura C.2 tiene la frecuencia fundamental $F_0 = \frac{1000}{8} = 125Hz$ pues el periodo es de 8 ms.

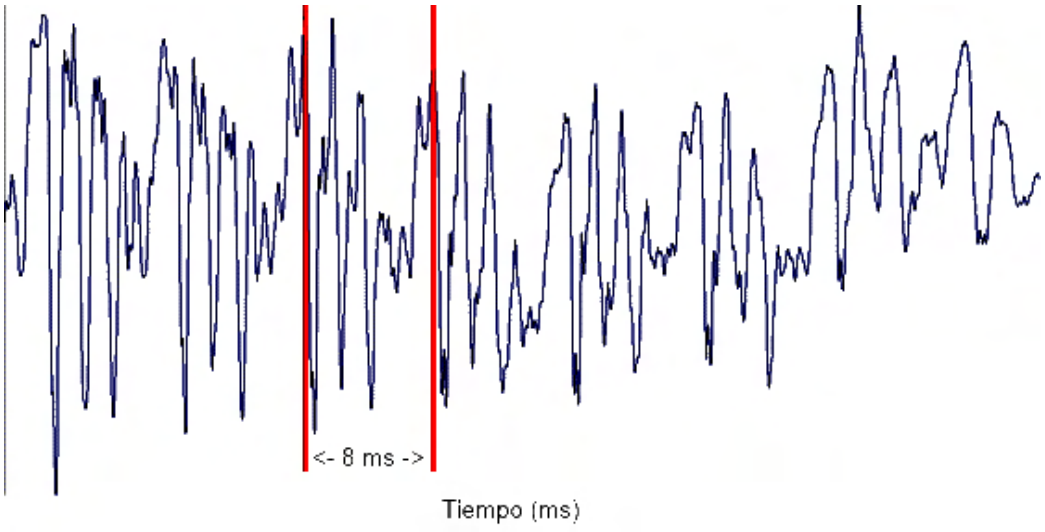

Figura C.2: *Forma de onda del sonido de una letra A*

C.1.2. Método de auto-correlación

Este método es muy similar al anterior, la señal es segmentada en ventanas $f_s(n : m)$ similares que serán comparadas hasta que se superpongan, es decir, hasta que sus diferencias sean casi 0. Como se puede ver en la Figura C.3.

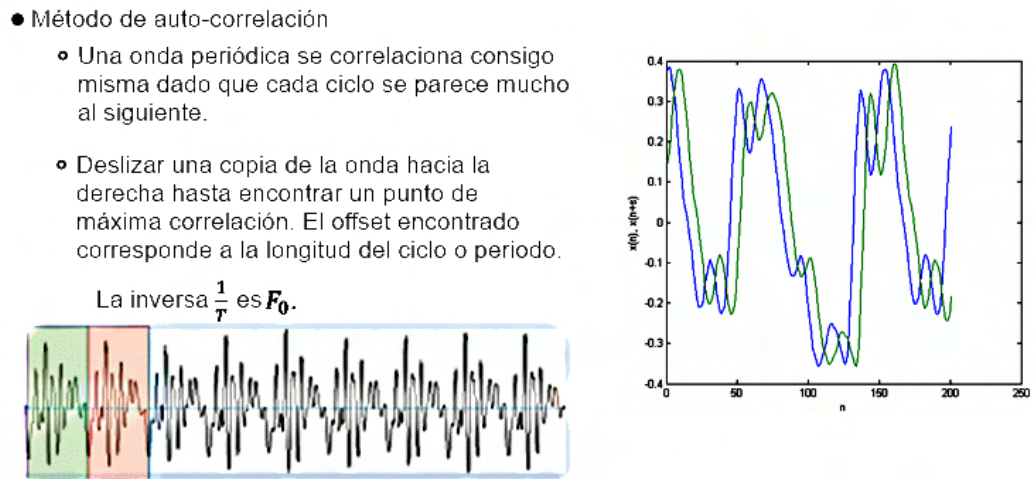

Figura C.3: *Auto-correlación*

Para esto, se filtran los tonos por debajo de 900 Hz y a cada ventana se le aplica un 'recorte' que impide que el primer formante interfiera con el pitch.

$$f_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & |f_s(n;m)| > C_{thr} \\ 0 & |f_s(n;m)| < C_{thr} \end{cases} \tag{C.2}$$

Donde C_{thr} es el 30 % del valor máximo en $f_s(n;m)$ para después calcular la auto-correlación.

$$r_{s(n;m)} = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} f_s(n;m) f_s(n-\eta;m) \tag{C.3}$$

Donde η es el offset. La frecuencia en la ventana m puede ser estimada con:

$$F_0(m) = \frac{F_s}{N_w} argmax_\eta |r(n;m) \begin{matrix} \eta = N_w \left(\frac{F_h}{F_s}\right) \\ \eta = N_w \left(\frac{F_l}{F_s}\right) \end{matrix} \tag{C.4}$$

Donde F_s es la frecuencia de muestreo y F_l y F_h son la menor y mayor frecuencia percibida por los humanos. Valores típicos para estos parámetros son F_s=8000 Hz, F_l=50 Hz y F_h=500 Hz. El valor máximo de la auto-correlación $argmax_\eta |r(n;m)_{\eta=N_w\left(\frac{F_l}{F_s}\right)}^{\eta=N_w\left(\frac{F_h}{F_s}\right)}$ será F_0.

C.1.3. Método de cepstrum

Hasta ahora hemos hablado del espectro y de la frecuencia, en este método se hablará de otros dos dominios relacionados con los anteriores:

- Cepstrum se derivó de Spectrum
- Quefrency se derivó de Frequency

El *liftering* es similar a la operación de filtrado en el dominio de la frecuencia donde se selecciona una región deseada para el análisis multiplicando el cepstrum por una ventana rectangular en la posición deseada. Hay dos tipos de liftering, bajo-tiempo y alto-tiempo:

1. De bajo-tiempo
    a) Para extraer las características del tracto vocal en el dominio de la cuefrencia.
2. De alto-tiempo
    a) Para obtener las características de excitación de la ventana de análisis.

Como se puede ver en la Figura C.4, el cepstrum es el resultado de tomar la transformada inversa discreta de Fourier (IDFT) del logaritmo de la magnitud del espectro de una señal. Es decir:

$$\begin{aligned} c(n) &= IDFT(\log|S(w)|) \\ &= IDFT(\log|E(w)| + \log|H(w)|) \end{aligned} \tag{C.5}$$

Figura C.4: *Método de cepstrum*

Si $e(n)$ es la secuencia de excitación y $h(n)$ es la secuencia del filtro del tracto vocal, entonces la secuencia de voz $s(n)$ se puede expresar de la siguiente manera:

$$\begin{aligned} s(n) &= e(n)h(n) \\ S(w) &= E(w)H(w) \end{aligned} \tag{C.6}$$

Tal que la magnitud del espectro de la señal de voz es:

$$|S(w)| = |E(w)||H(w)| \tag{C.7}$$

A fin de combinar linealmente ambos componentes en el dominio de la frecuencia, se aplica la representación logarítmica:

$$\log|S(w)| = \log|E(w)| + \log|H(w)| \tag{C.8}$$

Para la estimación de la frecuencia fundamental se aplica el *liftering* de alto-tiempo. Como el cepstrum calculado a partir de la secuencia del habla es simétrico, solo la mitad de la longitud del cepstrum se considera para el liftering. La característica de excitación se obtiene a través de una operación de alto-tiempo, utilizando una ventana rectangular (Figura C.5):

$$w_h(n) = \begin{cases} 1 & L_c \leq n < \frac{N}{2} \\ 0 & \end{cases} \tag{C.9}$$

O La ventana de Hamming (Figura C.5):

$$w_h(n) = 0{,}54 - 0{,}46\cos(\frac{2\pi n}{N}) \tag{C.10}$$

Donde L_c es la longitud de corte de la ventana de liftering y N es la longitud total del cepstrum. Por lo general se utiliza L_c como 15 ó 20.

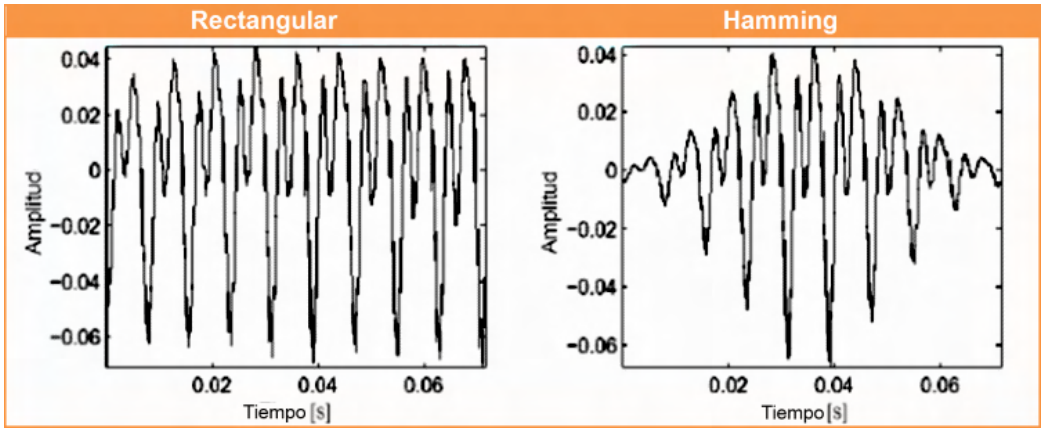

Figura C.5: *Tipo de ventanas*

Las características de excitación se obtienen multiplicando la ventana de liftering de alto-tiempo con el cepstrum obtenido de:

$$c_h(n) = w_h(n)c(n) \tag{C.11}$$

Así pues, el tono puede ser estimado como el instante correspondiente al pico más alto del lifter del cepstrum de alto-tiempo. Esto se puede ver esquematizado en la Figura C.6

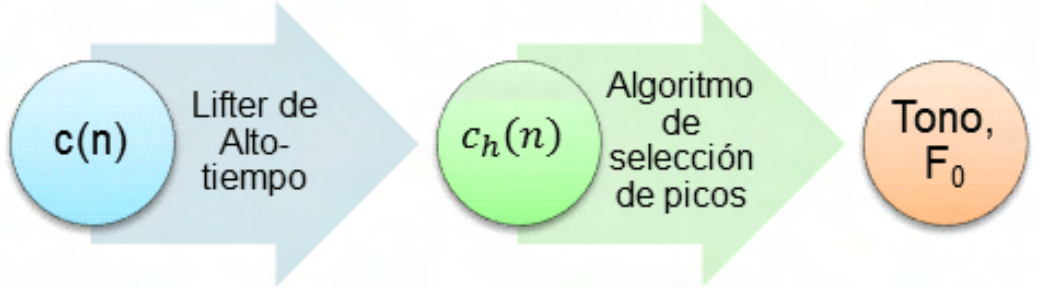

Figura C.6: *Algoritmo general para obtener F_0 mediante el cepstrum*

C.2. Coeficientes Cepstrales en la frecuencia de Mel

Los Mel frequency cepstral coefficients (MFCC) son coeficientes para la representación del habla basados en la percepción auditiva humana. Se derivan de la Transformada de Fourier y de la Transformada Discreta del Coseno, la diferencia radica en que en MFCC las bandas de frecuencia están situadas logarítmicamente según la escala Mel.

Para convertir de Hertz a Mels se utiliza la siguiente fórmula:

$$m = 2595 \log_1 01 + \frac{hz}{700} \tag{C.12}$$

Y, al contrario, para convertir de Mels a Hertz:

$$hz = 700(10^{\frac{m}{2595}} - 1) \tag{C.13}$$

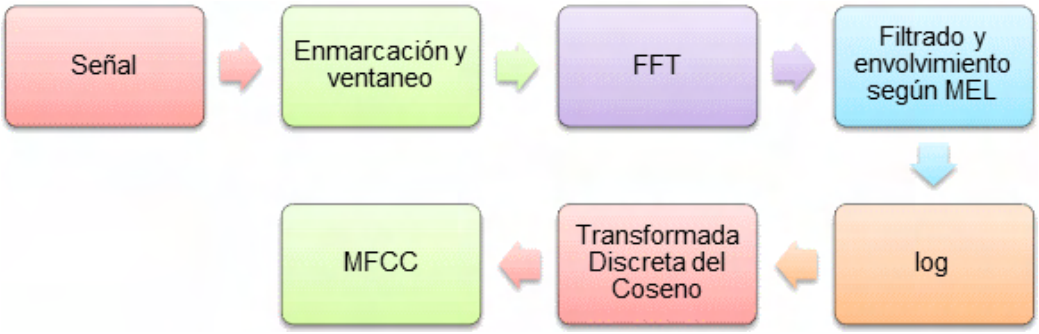

Figura C.7: *Obtención de los MFCCs*

Como está esquematizado en la Figura C.7, para obtener los MFCCs se toma la transformada de Fourier de un extracto de la ventana de una señal, se mapea la energía del espectro obtenido a la escala mel usando una función ventana triangular (Figura C.8), por ejemplo:

$$w(n) = \frac{2}{N+1}\left(\frac{N-1}{2} - \left|n - \frac{N-1}{2}\right|\right) \tag{C.14}$$

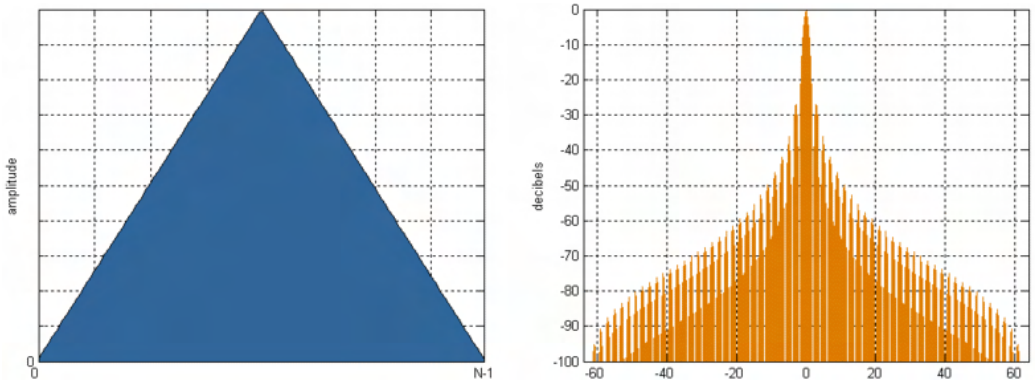

Figura C.8: *Ventana triangular*

Para después calcular el logaritmo de la energía de cada frecuencia Mel y como si fuera una señal, tomar la transformada de coseno discreta del resultado.

$$MFCC = FFT(Mel(\log(DCT(w)))) \tag{C.15}$$

Los MFCCs son las amplitudes del espectro resultante. Estas amplitudes se pueden ver en la Figura C.9.

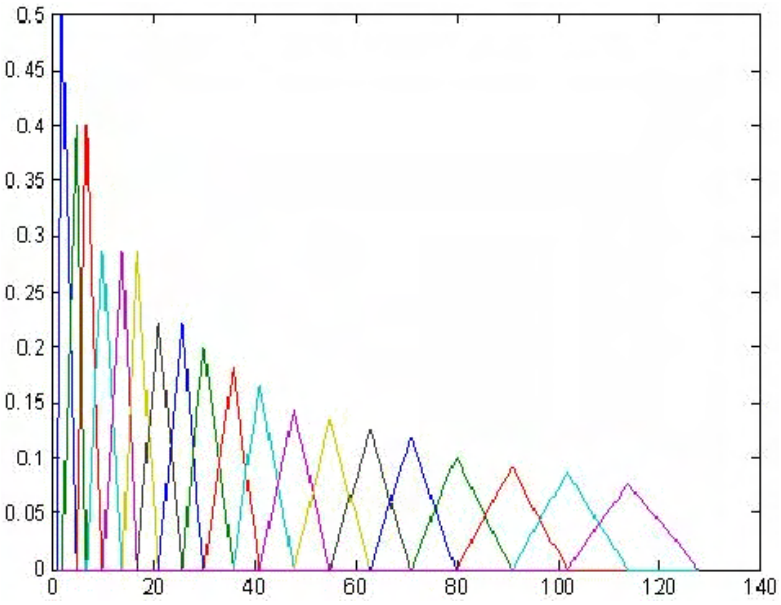

Figura C.9: *MFCCs*

C.3. Cruces por Cero

La Tasa de Cruces por Cero, es la tasa de cambios de signo en una señal como se puede ver en la Figura C.10, es decir, la tasa con la que la señal cambia de positivo a negativo o viceversa. Está definida como:

$$zcr = \frac{1}{T-1}\sum_{t=0}^{T-1} A(s_t s_{t-1}) \; con$$

$$A(s_t s_{t-1}) = \begin{cases} 1 & s_t s_{t-} < 0 \\ 0 & \end{cases} \tag{C.16}$$

Donde s es una señal de longitud T.

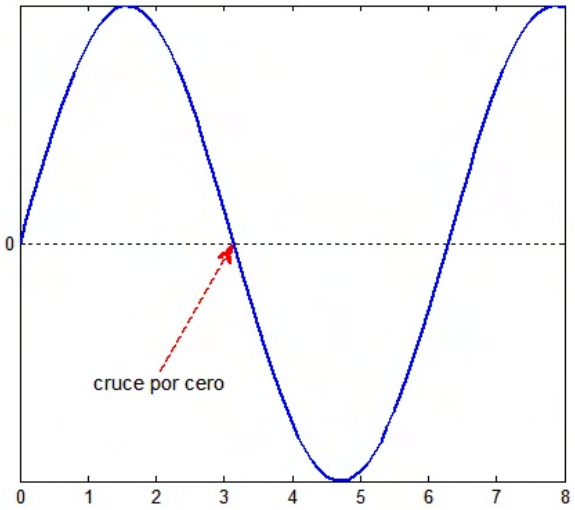

Figura C.10: *Cruce por cero*

C.4. Energía

Por naturaleza, la energía asociada con el habla es variable en el tiempo. De ahí el interés por saber cómo ésta característica está cambiando, más específicamente, cómo lo está haciendo en una región del segmento hablado, como en el caso de una palabra. Como sabemos, la señal de voz consiste en segmentos hablados y no hablados, es en estas regiones que la energía asociada a la presencia de la voz es grande en comparación con la región sin voz. Por lo tanto, la energía en segmentos cortos se puede utilizar para determinar aquellas regiones que nos sean de interés.

La relación para encontrar la energía a corto plazo se puede derivar de la relación de energía total. Ésta está dada por [74]:

$$E_T = \sum_{t=0}^{T-1} s^2(n) \tag{C.17}$$

Donde s es una señal de longitud T.

Para el caso del cálculo de energía en una ventana utilizamos la Ecuación C.1 como lo hemos venido haciendo en los puntos anteriores. En consecuencia se puede escribir la relación mencionada como:

$$e(n) = \sum_{t=0}^{T-1} [s(n)w(m-n)]^2 \tag{C.18}$$

Donde w puede ser cualquier tipo de ventana por ejemplo; rectangular definida por la Ecuación C.9, Hamming definida por la Ecuación C.10 o triangular definida por la Ecuación C.14.

APÉNDICE D

HEURÍSTICAS PARA EL ENTRENAMIENTO

Los parámetros que se modificaron durante el entrenamiento fueron: tamaño de la partición, que es el número de vectores de entrenamiento utilizados en cada pasada de cada época para el algoritmo de divergencia contrastiva, tasa de aprendizaje, número de unidades ocultas, y número de RBMs. Afortunadamente en [52] se presentan una serie de recomendaciones para elegir de manera informada estos parámetros. A continuación se describen las recomendaciones que fueron de utilidad en este trabajo.

Tamaño de la partición. Es posible actualizar los pesos después de estimar el gradiente en cada presentación de un vector de entrenamiento aunque a menudo es más eficiente dividir el conjunto de entrenamiento en pequeñas particiones de entre 10 y 100 vectores. Esto permite que las operaciones entre matrices se beneficien de Matlab o incluso del uso de GPU, por otro lado, es un error hacer las particiones muy grandes cuando se utiliza el gradiente descendente estocástico ya que, incrementar el tamaño de la partición en un factor de N no lo vuelve más fiable.

Para los conjuntos de datos que contienen un pequeño número de clases equiprobables, el tamaño ideal de la partición es a menudo igual al número de clases. Además, cada partición debe contener un ejemplo de cada clase para reducir el error de muestreo en la estimación del gradiente. Para los conjuntos de datos de otros tipos, conviene primero aleatorizar el orden de los ejemplos de entrenamiento y entonces utilizar particiones de aproximadamente tamaño 10.

Tasa de aprendizaje. Si la tasa de aprendizaje es muy grande, el error de reconstrucción aumentará drásticamente haciendo que los pesos puedan salirse de control. Si la tasa de aprendizaje se reduce mientras que la red está aprendiendo normalmente, el error de reconstrucción comenzará a decaer. Esto no es necesariamente bueno ya que en el largo plazo será de lento aprendizaje.

Una buena regla para la elección de la tasa de aprendizaje es ver un histograma de los cambios de peso y un histograma de los pesos. Las actualizaciones deben ser de aproximadamente 10^{-3} veces los pesos.

Número de unidades ocultas. En el aprendizaje discriminativo, la cantidad de restricción que impone un vector de entrenamiento es igual al número de bits que se necesita para representar la etiqueta. Las etiquetas suelen contener muy pocos bits de información, por lo que utilizar más parámetros que vectores de entrenamiento suele provocar un grave sobre entrenamiento. Sin embargo, cuando se está entrenando un modelo con datos de alta dimensionalidad, es el número de bits que se necesita para representar un vector de entrenamiento lo que determina la restricción.

Suponiendo que el tema principal es el sobre entrenamiento, se debe estimar la cantidad de bits que se necesitaría para describir cada vector de datos, luego se multiplica ese estimado por el número de vectores

de entrenamiento. El resultado se utiliza con un orden de magnitud menor. Si los datos sobre los que se está entrenando son pocos, puede que sea buena idea utilizar más unidades ocultas. Por el contrario, si los datos de entrenamiento son altamente redundantes, entonces será buena idea utilizar menos unidades ocultas.

APÉNDICE E

TABLAS DE RESULTADOS

Los resultados de los experimentos presentados en la Tabla 6.1, donde se utilizó una RBM como clasificador, se muestran a continuación. La combinación de parámetros con los que se obtuvieron los mejores resultados fueron:

- Tamaño de partición: 54
- Tasa de aprendizaje: 0.0001
- Unidades ocultas: 84

Con esta configuración, la DBN con una RBM alcanzó una tasa de error de 2,51 %.

Sub experimento			% Error						
T. partición	T. aprendizaje	U. ocultas	DBN	KNN	RBM KNN	DTree	RBM DTree	MLP	RBM MLP
6	0.01	28	68.1	2.87	54.84	11.83	15.05	5.15	51.61
6	0.01	56	50.9	2.87	49.46	9.68	6.45	4.07	36.56
6	0.01	84	44.45	2.87	66.67	9.68	9.68	5.15	59.14
6	0.01	112	45.52	2.87	67.74	9.68	10.75	5.15	38.71
6	0.01	140	77.78	2.87	56.99	9.68	5.38	68.59	60.22
6	0.01	168	55.2	2.87	66.67	9.68	6.45	4.07	54.84
6	0.001	28	5.74	2.87	54.84	9.68	6.45	4.07	58.06
6	0.001	56	15.41	2.87	44.09	9.68	11.83	4.07	39.78
6	0.001	84	10.04	2.87	37.63	9.68	16.13	4.07	37.63
6	0.001	112	15.41	2.87	66.67	9.68	7.53	4.07	58.06
6	0.001	140	10.04	2.87	68.82	9.68	8.6	3	56.99
6	0.001	168	4.66	2.87	66.67	9.68	11.83	35.25	59.14
6	0.0001	28	5.74	2.87	66.67	9.68	6.45	5.15	60.22
6	0.0001	56	6.81	2.87	67.74	9.68	11.83	4.07	58.06
6	0.0001	84	7.89	2.87	55.91	11.83	7.53	6.22	55.91
6	0.0001	112	6.81	2.87	34.41	9.68	8.6	4.07	66.67
6	0.0001	140	8.96	2.87	66.67	9.68	13.98	4.07	54.84
6	0.0001	168	10.04	2.87	39.78	9.68	10.75	5.15	41.94
6	0.00001	28	5.74	2.87	37.63	9.68	5.38	6.22	46.24
6	0.00001	56	4.66	2.87	55.91	9.68	7.53	5.15	52.69
6	0.00001	84	8.96	2.87	10.75	9.68	3.23	7.3	9.68
6	0.00001	112	4.66	2.87	11.83	11.83	5.38	3	12.9
6	0.00001	140	7.89	2.87	19.35	9.68	8.6	5.15	51.61
6	0.00001	168	7.89	2.87	6.45	9.68	4.3	5.15	8.6
12	0.01	28	65.59	2.87	7.53	9.68	8.6	5.15	9.68
12	0.01	56	48.39	2.87	33.33	9.68	8.6	3	38.71
12	0.01	84	8.96	2.87	7.53	9.68	4.3	36.33	10.75
12	0.01	112	15.41	2.87	7.53	9.68	7.53	5.15	10.75
12	0.01	140	75.27	2.87	8.6	9.68	6.45	6.22	7.53
12	0.01	168	52.69	2.87	8.6	11.83	5.38	4.07	9.68
12	0.001	28	3.59	2.87	8.6	9.68	6.45	37.4	12.9
12	0.001	56	12.19	2.87	4.3	11.83	5.38	5.15	6.45
12	0.001	84	5.74	2.87	8.6	9.68	4.3	4.07	10.75
12	0.001	112	6.81	2.87	8.6	9.68	8.6	5.15	38.71

12	0.001	140	2.51	2.87	6.45	9.68	6.45	6.22	11.83
12	0.001	168	5.74	2.87	9.68	9.68	5.38	4.07	12.9
12	0.0001	28	5.74	2.87	17.2	9.68	5.38	5.15	11.83
12	0.0001	56	4.66	2.87	8.6	9.68	5.38	4.07	16.13
12	0.0001	84	4.66	2.87	13.98	9.68	7.53	4.07	12.9
12	0.0001	112	3.59	2.87	5.38	11.83	7.53	4.07	4.3
12	0.0001	140	5.74	2.87	23.66	9.68	13.98	3	18.28
12	0.0001	168	4.66	2.87	9.68	9.68	4.3	5.15	39.78
12	0.00001	28	6.81	2.87	12.9	9.68	8.6	4.07	18.28
12	0.00001	56	3.59	2.87	7.53	9.68	7.53	4.07	8.6
12	0.00001	84	2.51	2.87	8.6	9.68	4.3	4.07	10.75
12	0.00001	112	4.66	2.87	8.6	9.68	8.6	5.15	7.53
12	0.00001	140	4.66	2.87	8.6	11.83	6.45	4.07	4.3
12	0.00001	168	2.51	2.87	7.53	9.68	3.23	4.07	6.45
18	0.01	28	3.59	2.87	37.63	9.68	11.83	4.07	26.88
18	0.01	56	4.66	2.87	67.74	9.68	8.6	68.59	63.44
18	0.01	84	16.49	2.87	30.11	9.68	9.68	4.07	52.69
18	0.01	112	11.11	2.87	66.67	11.83	6.45	6.22	66.67
18	0.01	140	4.66	2.87	25.81	10.75	10.75	6.22	41.94
18	0.01	168	3.59	2.87	66.67	9.68	9.68	35.25	56.99
18	0.001	28	8.96	2.87	51.61	9.68	8.6	5.15	56.99
18	0.001	56	4.66	2.87	50.54	9.68	15.05	6.22	53.76
18	0.001	84	4.66	2.87	31.18	9.68	13.98	5.15	36.56
18	0.001	112	6.81	2.87	68.82	9.68	4.3	5.15	59.14
18	0.001	140	7.89	2.87	68.82	9.68	8.6	6.22	61.29
18	0.001	168	5.74	2.87	6.45	9.68	4.3	6.22	6.45
18	0.0001	28	35.84	2.87	9.68	9.68	4.3	35.25	9.68
18	0.0001	56	4.66	2.87	70.97	9.68	6.45	4.07	58.06
18	0.0001	84	7.89	2.87	43.01	9.68	8.6	5.15	56.99
18	0.0001	112	2.51	2.87	27.96	9.68	9.68	5.15	64.52
18	0.0001	140	7.89	2.87	5.38	9.68	6.45	4.07	8.6
18	0.0001	168	4.66	2.87	9.68	9.68	5.38	5.15	12.9
18	0.00001	28	6.81	2.87	8.6	9.68	6.45	5.15	9.68
18	0.00001	56	5.74	2.87	7.53	9.68	3.23	4.07	6.45
18	0.00001	84	5.74	2.87	9.68	9.68	4.3	5.15	10.75
18	0.00001	112	3.59	2.87	7.53	11.83	6.45	4.07	12.9
18	0.00001	140	2.51	2.87	7.53	9.68	3.23	3	8.6
18	0.00001	168	2.51	2.87	7.53	9.68	4.3	3	6.45
24	0.01	28	53.76	2.87	5.38	9.68	4.3	4.07	7.53
24	0.01	56	5.74	2.87	3.23	9.68	5.38	5.15	9.68

24	0.01	84	3.59	2.87	6.45	10.75	4.3	4.07	6.45
24	0.01	112	10.04	2.87	7.53	9.68	6.45	4.07	6.45
24	0.01	140	5.74	2.87	7.53	9.68	6.45	5.15	6.45
24	0.01	168	3.58	2.87	5.38	9.68	11.83	6.22	6.45
24	0.001	28	6.81	2.87	8.6	9.68	4.3	6.22	5.38
24	0.001	56	4.66	2.87	5.38	11.83	6.45	5.15	5.38
24	0.001	84	5.74	2.87	50.54	9.68	20.43	3	55.91
24	0.001	112	11.11	2.87	8.6	9.68	3.23	35.25	10.75
24	0.001	140	2.51	2.87	6.45	9.68	8.6	7.3	5.38
24	0.001	168	4.66	2.87	5.38	11.83	7.53	38.48	9.68
24	0.0001	28	4.66	2.87	6.45	9.68	7.53	5.15	7.53
24	0.0001	56	3.59	2.87	3.23	9.68	3.23	4.07	3.23
24	0.0001	84	2.51	2.87	5.38	11.83	4.3	3	4.3
24	0.0001	112	3.59	2.87	6.45	9.68	7.53	4.07	4.3
24	0.0001	140	6.81	2.87	7.53	9.68	6.45	33.1	8.6
24	0.0001	168	10.04	2.87	3.23	11.83	5.38	5.15	5.38
24	0.00001	28	3.59	2.87	5.38	9.68	7.53	5.15	7.53
24	0.00001	56	7.89	2.87	12.9	9.68	7.53	6.22	13.98
24	0.00001	84	5.74	2.87	6.45	9.68	3.23	7.3	8.6
24	0.00001	112	4.66	2.87	8.6	9.68	8.6	6.22	5.38
24	0.00001	140	3.59	2.87	6.45	9.68	4.3	4.07	6.45
24	0.00001	168	4.66	2.87	6.45	9.68	4.3	4.07	66.67
30	0.01	28	18.64	2.87	67.74	9.68	13.98	68.59	36.56
30	0.01	56	10.04	2.87	34.41	11.83	7.53	5.15	38.71
30	0.01	84	5.74	2.87	13.98	9.68	18.28	4.07	17.2
30	0.01	112	11.11	2.87	21.51	11.83	5.38	6.22	61.29
30	0.01	140	12.19	2.87	12.9	11.83	6.45	4.07	15.05
30	0.01	168	2.51	2.87	12.9	9.68	6.45	5.15	10.75
30	0.001	28	5.74	2.87	67.74	9.68	6.45	4.07	55.91
30	0.001	56	4.66	2.87	66.67	11.83	58.06	4.07	66.67
30	0.001	84	2.51	2.87	10.75	9.68	5.38	4.07	9.68
30	0.001	112	5.74	2.87	22.58	9.68	9.68	5.15	22.58
30	0.001	140	5.74	2.87	4.3	11.83	5.38	5.15	9.68
30	0.001	168	4.66	2.87	11.83	9.68	5.38	6.22	20.43
30	0.0001	28	4.66	2.87	6.45	11.83	4.3	4.07	16.13
30	0.0001	56	5.74	2.87	5.38	11.83	4.3	4.07	9.68
30	0.0001	84	3.59	2.87	10.75	9.68	4.3	5.15	13.98
30	0.0001	112	3.59	2.87	9.68	9.68	3.23	6.22	12.9
30	0.0001	140	5.74	2.87	67.74	9.68	7.53	5.15	56.99
30	0.0001	168	3.59	2.87	5.38	9.68	3.23	5.15	7.53

30	0.00001	28	6.81	2.87	7.53	9.68	8.6	5.15	8.6
30	0.00001	56	3.59	2.87	10.75	9.68	11.83	5.15	10.75
30	0.00001	84	3.59	2.87	7.53	9.68	8.6	7.3	8.6
30	0.00001	112	3.59	2.87	7.53	9.68	7.53	5.15	6.45
30	0.00001	140	3.59	2.87	6.45	11.83	3.23	68.59	8.6
30	0.00001	168	3.59	2.87	6.45	9.68	5.38	5.15	4.3
36	0.01	28	7.89	2.87	7.53	9.68	5.38	6.22	13.98
36	0.01	56	10.04	2.87	29.03	9.68	8.6	5.15	52.69
36	0.01	84	4.66	2.87	5.38	9.68	7.53	35.25	31.18
36	0.01	112	7.89	2.87	17.2	9.68	8.6	4.07	16.13
36	0.01	140	3.59	2.87	5.38	9.68	6.45	37.4	10.75
36	0.01	168	6.81	2.87	8.6	9.68	5.38	4.07	5.38
36	0.001	28	6.81	2.87	5.38	9.68	4.3	6.22	3.23
36	0.001	56	15.41	2.87	6.45	9.68	9.68	5.15	5.38
36	0.001	84	6.81	2.87	19.35	9.68	7.53	4.07	15.05
36	0.001	112	5.74	2.87	7.53	9.68	5.38	6.22	9.68
36	0.001	140	5.74	2.87	5.38	9.68	7.53	3	9.68
36	0.001	168	5.74	2.87	9.68	9.68	3.23	4.07	4.3
36	0.0001	28	4.66	2.87	5.38	9.68	3.23	4.07	10.75
36	0.0001	56	2.51	2.87	8.6	11.83	6.45	4.07	7.53
36	0.0001	84	3.59	2.87	8.6	9.68	3.23	4.07	4.3
36	0.0001	112	2.51	2.87	7.53	9.68	3.23	5.15	5.38
36	0.0001	140	4.66	2.87	8.6	9.68	5.38	3	8.6
36	0.0001	168	5.74	2.87	6.45	11.83	4.3	4.07	5.38
36	0.00001	28	3.59	2.87	8.6	11.83	5.38	68.59	8.6
36	0.00001	56	5.74	2.87	3.23	9.68	5.38	5.15	6.45
36	0.00001	84	4.66	2.87	6.45	9.68	6.45	35.25	4.3
36	0.00001	112	2.51	2.87	6.45	9.68	5.38	5.15	36.56
36	0.00001	140	3.58	2.87	4.3	9.68	3.23	35.25	5.38
36	0.00001	168	3.59	2.87	4.3	9.68	7.53	5.15	4.3
42	0.01	28	7.89	2.87	5.38	9.68	7.51	5.15	16.13
42	0.01	56	4.66	2.87	6.45	9.68	4.3	3	7.53
42	0.01	84	5.74	2.87	5.38	9.68	5.38	4.07	33.33
42	0.01	112	2.51	2.87	11.83	9.68	4.3	4.07	8.6
42	0.01	140	7.89	2.87	11.83	11.83	7.53	5.15	9.68
42	0.01	168	5.74	2.87	5.38	9.68	4.3	4.07	8.6
42	0.001	28	3.59	2.87	6.45	9.68	6.45	6.22	5.38
42	0.001	56	5.74	2.87	66.67	9.68	51.61	5.15	51.61
42	0.001	84	3.59	2.87	20.43	9.68	6.45	4.07	23.66
42	0.001	112	5.74	2.87	8.6	9.68	5.38	4.07	7.53

42	0.001	140	2.51	2.87	7.53	9.68	5.38	5.15	5.38
42	0.001	168	4.66	2.87	6.45	9.68	7.53	5.15	66.67
42	0.0001	28	4.66	2.87	6.45	9.68	7.53	5.15	7.53
42	0.0001	56	3.59	2.87	7.53	11.83	5.38	3	8.6
42	0.0001	84	5.74	2.87	4.3	9.68	9.68	6.22	9.68
42	0.0001	112	8.96	2.87	66.67	9.68	51.61	4.07	51.61
42	0.0001	140	3.59	2.87	5.38	9.68	4.3	4.07	8.6
42	0.0001	168	3.59	2.87	8.6	9.68	3.23	3	8.6
42	0.00001	28	3.59	2.87	5.38	9.68	7.53	4.07	7.53
42	0.00001	56	7.89	2.87	7.53	9.68	3.23	4.07	5.38
42	0.00001	84	6.81	2.87	7.53	9.68	5.38	4.07	8.6
42	0.00001	112	4.66	2.87	5.38	11.83	4.3	5.15	4.3
42	0.00001	140	5.74	2.87	7.53	9.68	3.23	5.15	7.53
42	0.00001	168	8.96	2.87	4.3	9.68	11.83	7.3	6.45
48	0.01	28	11.11	2.87	72.04	9.68	15.05	5.15	61.29
48	0.01	56	10.04	2.87	4.3	9.68	3.23	4.07	37.63
48	0.01	84	10.04	2.87	4.3	11.83	8.6	4.07	9.68
48	0.01	112	3.59	2.87	3.23	9.68	5.38	4.07	6.45
48	0.01	140	7.89	2.87	5.38	9.68	4.3	6.22	6.45
48	0.01	168	5.74	2.87	8.6	9.68	6.45	5.15	5.38
48	0.001	28	5.74	2.87	6.45	9.68	5.38	6.22	4.3
48	0.001	56	4.66	2.87	7.53	9.68	4.3	4.07	25.81
48	0.001	84	5.74	2.87	8.6	9.68	6.45	6.22	11.83
48	0.001	112	3.59	2.87	8.6	9.68	6.45	4.07	8.6
48	0.001	140	7.89	2.87	15.05	9.68	8.6	6.22	12.9
48	0.001	168	2.51	2.87	8.6	11.83	3.23	5.15	7.53
48	0.0001	28	7.89	2.87	8.6	9.68	5.38	35.25	7.53
48	0.0001	56	7.89	2.87	8.6	9.68	3.23	5.15	4.3
48	0.0001	84	5.74	2.87	7.53	11.83	4.3	5.15	6.45
48	0.0001	112	4.66	2.87	6.45	9.68	4.3	4.07	3.23
48	0.0001	140	4.66	2.87	19.35	9.68	16.13	6.22	38.71
48	0.0001	168	2.51	2.87	7.53	11.83	4.3	5.15	7.53
48	0.00001	28	3.59	2.87	5.38	9.68	5.38	5.15	7.53
48	0.00001	56	4.66	2.87	7.53	9.68	6.45	3	9.68
48	0.00001	84	4.66	2.87	5.38	11.83	5.38	6.22	6.45
48	0.00001	112	4.66	2.87	4.3	9.68	9.68	35.25	5.38
48	0.00001	140	3.59	2.87	6.45	9.68	4.3	4.07	5.38
48	0.00001	168	3.59	2.87	5.38	9.68	7.53	5.15	5.38
54	0.01	28	4.66	2.87	56.99	9.68	17.2	4.07	66.67
54	0.01	56	8.96	2.87	7.53	9.68	3.23	5.15	8.6

54	0.01	84	5.74	2.87	7.53	9.68	10.75	3	9.68
54	0.01	112	6.81	2.87	5.38	9.68	5.38	4.07	5.38
54	0.01	140	7.89	2.87	4.3	9.68	6.45	5.15	37.63
54	0.01	168	8.96	2.87	24.73	9.68	7.53	5.15	37.63
54	0.001	28	4.66	2.87	6.45	9.68	6.45	5.15	6.45
54	0.001	56	3.59	2.87	66.67	9.68	20.43	5.15	66.67
54	0.001	84	3.59	2.87	10.75	11.83	6.45	39.55	6.45
54	0.001	112	5.74	2.87	8.6	11.83	4.3	5.15	6.45
54	0.001	140	5.74	2.87	6.45	9.68	3.23	7.3	4.3
54	0.001	168	10.04	2.87	5.38	9.68	7.53	6.22	4.3
54	0.0001	28	3.59	2.87	4.3	9.68	3.23	3	4.3
54	0.0001	56	5.74	2.87	8.6	9.68	6.45	3	12.9
54	0.0001	84	2.51	2.87	2.51	6.45	5.38	3.58	3.51
54	0.0001	112	10.04	2.87	7.53	9.68	3.23	4.07	5.38
54	0.0001	140	5.74	2.87	10.75	9.68	4.3	6.22	7.53
54	0.0001	168	5.74	2.87	4.3	9.68	4.3	4.07	4.3
54	0.00001	28	8.96	2.87	5.38	9.68	4.3	4.07	7.53
54	0.00001	56	3.59	2.87	5.38	9.68	8.6	4.07	4.3
54	0.00001	84	4.66	2.87	6.45	9.68	5.23	4.07	8.6
54	0.00001	112	4.66	2.87	5.38	11.83	6.45	4.07	6.45
54	0.00001	140	3.59	2.87	3.23	9.68	5.38	5.15	37.63
54	0.00001	168	3.59	2.87	7.53	9.68	6.45	6.22	3.23
60	0.01	28	5.74	2.87	7.53	9.68	4.3	6.22	9.68
60	0.01	56	7.89	2.87	4.3	9.68	6.45	5.15	7.53
60	0.01	84	3.59	2.87	4.3	9.68	8.51	4.07	4.3
60	0.01	112	7.89	2.87	7.53	9.68	6.45	4.07	4.3
60	0.01	140	6.81	2.87	3.23	11.83	4.3	5.15	5.38
60	0.01	168	2.51	2.87	7.53	9.68	7.53	4.07	5.38
60	0.001	28	7.89	2.87	7.53	9.68	6.45	4.07	5.38
60	0.001	56	4.66	2.87	8.6	9.68	6.45	6.22	8.6
60	0.001	84	3.59	2.87	8.6	9.68	6.45	4.07	8.6
60	0.001	112	3.59	2.87	4.3	9.68	5.38	4.07	6.45
60	0.001	140	5.74	2.87	5.38	9.68	7.53	4.07	4.3
60	0.001	168	3.59	2.87	8.6	9.68	6.45	6.22	7.53
60	0.0001	28	5.74	2.87	6.45	9.68	4.3	5.15	7.53
60	0.0001	56	2.51	2.87	7.53	9.68	6.45	4.07	3.23
60	0.0001	84	7.89	2.87	7.53	9.68	10.75	3	7.53
60	0.0001	112	4.66	2.87	5.38	9.68	10.23	4.07	5.38
60	0.0001	140	4.66	2.87	31.18	9.68	17.2	5.15	35.48
60	0.0001	168	7.89	2.87	4.3	9.68	4.3	5.15	7.53

60	0.00001	28	6.81	2.87	8.6	9.68	5.38	4.07	4.3
60	0.00001	56	4.66	2.87	3.23	9.68	3.23	4.07	4.3
60	0.00001	84	5.74	2.87	6.45	9.68	4.3	38.48	3.23
60	0.00001	112	4.66	2.87	4.3	9.68	3.23	5.15	4.3
60	0.00001	140	8.96	2.87	8.6	9.68	4.3	5.15	8.6
60	0.00001	168	10.04	2.87	6.45	9.68	3.23	5.15	3.23