Desarrollo de un sistema de reconocimiento de emociones en el habla

Por

María Belén Crolla

DIRECTOR DEL PROYECTO
Dr. Diego H. Milone

CO-DIRECTOR DEL PROYECTO
Ing. Enrique M. Albornoz

PROYECTO FINAL DE CARRERA
INGENIERÍA EN INFORMÁTICA

Facultad de Ingeniería y Ciencias Hídricas
UNIVERSIDAD NACIONAL DEL LITORAL

1 de marzo de 2010

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. B. Crolla, D. H. Milone & E. M. Albornoz; "Desarrollo de un sistema de reconocimiento de emociones en el habla (Undergraduate project)"
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral, 2010.

Índice general

1. Introducción **1**
- 1.1. Motivación . . . 1
  - 1.1.1. Significado de la palabra emoción . . . 2
  - 1.1.2. La voz de las emociones . . . 3
  - 1.1.3. Papel de las emociones en la interacción hombre - computadora . . . 3
- 1.2. Antecedentes . . . 4
- 1.3. Aplicaciones del reconocedor de emociones . . . 7
- 1.4. Objetivos del proyecto . . . 8

2. Extracción de características y métodos de clasificación **10**
- 2.1. Análisis de señales de voz . . . 10
- 2.2. Mezcla de gaussianas . . . 13
- 2.3. Modelos ocultos de Markov . . . 15
  - 2.3.1. Problema de evaluación . . . 19
  - 2.3.2. Problema de estimación . . . 24
  - 2.3.3. Problema de entrenamiento . . . 26
  - 2.3.4. Modelo de lenguaje . . . 28

3. Corpus de emociones **29**
- 3.1. ¿Emociones actuadas o reales? . . . 30
- 3.2. Base de datos de habla emocional en alemán . . . 30
  - 3.2.1. Generalidades . . . 30
  - 3.2.2. La elección de los actores . . . 31
  - 3.2.3. Grabación de los audios . . . 31
  - 3.2.4. Test perceptual . . . 32
- 3.3. Base de datos de habla emocional en español . . . 33
  - 3.3.1. Generalidades . . . 33
  - 3.3.2. La elección de los actores . . . 34

3.3.3. Extracción de los segmentos de audio 35
3.3.4. Test perceptual 35

4. Diseño e implementación del sistema **37**
4.1. Extracción de características 37
4.2. Definición del modelo de referencia 38
4.3. Entrenamiento del modelo 40
4.4. Prueba del modelo 41
4.5. Alternativas de diseño evaluadas 42

5. Evaluación de resultados **44**
5.1. Resultados con el corpus de emociones en alemán 45
5.1.1. Reconocimiento de emociones con GMM 45
5.1.2. Reconocimiento de emociones con HMM 45
5.2. Resultados con el corpus de emociones en español 50
5.2.1. Reconocimiento de emociones con GMM 51
5.2.2. Reconocimiento de emociones con HMM 51

6. Conclusiones y trabajos futuros **56**

A. Corpus en alemán **58**

B. Corpus en español **59**

Índice de tablas

3.1. Distribución del corpus de emociones en alemán. 31
3.2. Distribución de los audios extraídos de películas en español. . 34
3.3. Distribución del corpus de emociones en español utilizado en el reconocedor de emociones. 35

5.1. Matriz de confusión para 3 emociones con GMM (22 gaussianas)[BD en alemán] 46
5.2. Matriz de confusión para 4 emociones con GMM (26 gaussianas)[BD en alemán] 46
5.3. Matriz de confusión para 5 emociones con GMM (32 gaussianas)[BD en alemán] 46
5.4. Matriz de confusión para 6 emociones con GMM (16 gaussianas)[BD en alemán] 46
5.5. Matriz de confusión para 7 emociones con GMM (32 gaussianas)[BD en alemán] 46
5.6. Matriz de confusión para 3 emociones con HMM (2 estados, 14 gaussianas)[BD en alemán] 49
5.7. Matriz de confusión para 4 emociones con HMM (2 estados, 30 gaussianas)[BD en alemán] 49
5.8. Matriz de confusión para 5 emociones con HMM (2 estados, 32 gaussianas)[BD en alemán] 49
5.9. Matriz de confusión para 6 emociones con HMM (2 estados, 24 gaussianas)[BD en alemán] 49
5.10. Matriz de confusión para 7 emociones con HMM (2 estados, 30 gaussianas)[BD en alemán] 49
5.11. Matriz de confusión para 3 emociones con GMM (18 gaussianas)[BD en español] 51
5.12. Matriz de confusión para 2 estados con HMM (14 gaussianas)[BD en español] 53

5.13. Matriz de confusión para 3 estados con HMM (32 gaussianas)[BD en español] 53
5.14. Matriz de confusión para 4 estados con HMM (28 gaussianas)[BD en español] 54
5.15. Matriz de confusión para 5 estados con HMM (4 gaussianas)[BD en español] 54
5.16. Matriz de confusión para 6 estados con HMM (6 gaussianas)[BD en español] 54
5.17. Matriz de confusión para 7 estados con HMM (26 gaussianas)[BD en español] 54

Índice de figuras

2.1. Distribución modelada por la mezcla de 3 gaussianas simples en una dimensión. 15
2.2. Diagrama de un modelo oculto de Markov de 5 estados. . . . 16
2.3. Cálculo de la variable hacia adelante. 21
2.4. Cálculo de la variable hacia atrás. 23
2.5. Modelo de lenguaje 28

3.1. Un actor durante la grabación en la cámara anecoica. 32
3.2. Resultados del test de percepción con la base de datos en alemán. 33
3.3. Resultados del test de percepción con la base de datos en español. 36

4.1. Red gramatical 39
4.2. Diagrama de bloques: fase de entrenamiento 40
4.3. Entrenamiento de un GMM/HMM 41
4.4. Diagrama de bloques: fase de reconocimiento 42

5.1. Reconocimiento de emociones en función de la cantidad de estados utilizando HMM [BD en alemán] 47
5.2. Reconocimiento de emociones en función de la cantidad de gaussianas, con modelos HMM de 2 estados [BD en alemán] . 48
5.3. Reconocimiento de emociones en función de la cantidad de gaussianas con GMM [BD en español] 52
5.4. Reconocimiento de emociones en función de la cantidad de estados utilizando HMM [BD en español] 53

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. B. Crolla, D. H. Milone & E. M. Albornoz; "Desarrollo de un sistema de reconocimiento de emociones en el habla (Undergraduate project)"
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral, 2010.

Prefacio

El reconocimiento de emociones es un área que ha cobrado gran importancia en los últimos años. La nueva tendencia en interacción hombre-máquina ha despertado gran inquietud en el campo de las expresiones emocionales y el fin es lograr una relación más natural y amigable con el usuario. La falta de información en estas interacciones puede ser crucial a la hora de tomar decisiones por parte de la máquina. En muchas ocasiones la interacción falla debido a que el humano experimenta la falta de capacidad, por parte del dispositivo, de interactuar de modo similar a un humano.

En este trabajo se presentan dos métodos capaces de reconocer de forma automática la emoción que expresa la persona a partir de su registro de voz. Esta información es importante ya que permite que el dispositivo tome decisiones de acuerdo a la emoción expresada por el usuario. Resulta de gran utilidad que el reconocimiento sea a través de la voz, ya que en la misma está contenida la información explícita e implícita del mensaje que el locutor quiere expresar.

Este proyecto final está estructurado en seis capítulos y su orden se corresponde con las etapas desarrolladas. En el primer capítulo, se introducen los conceptos y las formas de clasificar las emociones así como también su importancia en una interfase hombre-máquina. Aquí también se plantea la motivación del proyecto final de carrera, los antecedentes de investigaciones en este campo y los objetivos que se persiguen. En el segundo capítulo se trata el marco teórico necesario para el desarrollo de un sistema de reconocimiento de emociones. Se describe el análisis de las señales de voz y se explican los métodos de clasificación: los modelos de mezclas de gaussianas y los modelos ocultos de Markov. En el tercer capítulo se desarrollan temas referidos a la creación y validación de las bases de datos de emociones utilizadas, tanto para el alemán como para el español. En el cuarto capítulo se describe detalladamente los pasos seguidos en el desarrollo del reconocedor y los diferentes diseños evaluados para obtener una mejor tasa de reconocimiento. En el quinto capítulo, se exponen y discuten todos los resultados

obtenidos con los diferentes métodos y con las dos bases de datos. En el capítulo final, se presentan las conclusiones obtenidas durante el desarrollo del proyecto y se proponen posibles líneas futuras de trabajo.

Capítulo 1

Introducción

En este capítulo se introducen conceptos de diferentes disciplinas que intervienen en el desarrollo de este trabajo. Principalmente se revisa el concepto de emoción, así como su clasificación y representación, ya que es uno de los factores más íntimamente relacionados con la expresión humana. A continuación, se introducen los antecedentes en el campo de investigación de las emociones. Por último, se describen los objetivos propuestos para el desarrollo de este proyecto.

1.1. Motivación

Hoy en día, con el progreso de las nuevas tecnologías, es muy común la interacción hombre-máquina. Debido a que los humanos tienden a expresar el estado emocional a través de la voz, que las máquinas logren la capacidad de reconocer automáticamente los estados emocionales, es menester para mejorar la interacción. Los sistemas actuales de interacción hombre-máquina basados en voz pueden reconocer "que dijo" y "quien lo dijo" utilizando técnicas de reconocimiento del habla y de identificación del hablante. Si se adicionara a estos sistemas el reconocimiento de emociones, podrían saber "en que estado emotivo fue dicho" para actuar en consecuencia logrando una interacción más natural. La comunicación humana se realiza por dos canales diferentes [1]: uno se encarga de la transmisión de los mensajes de forma explícita mientras que el otro lo hace de forma implícita, aportando información sobre el propio locutor. Es en este segundo canal, no tan analizado actualmente, en el cual se centra el estudio del reconocimiento de emociones [2]. Aunque se continúa trabajando en el ámbito del canal explícito, cada vez cobran mayor importancia las investigaciones sobre el canal implícito inten-

tando conseguir una mejor interacción hombre-máquina con interfaces que emulen mejor el proceso de reconocimiento. Es muy importante estudiar el reconocimiento de emociones y añadirlo a una interfaz automática. Algunos autores del campo de la psicología [3, 4] indican que este tipo de reconocimiento por parte de los interlocutores (también conocido como *empatía*) es la base de las relaciones humanas, y se fundamenta en la interpretación de las señales transmitidas de forma inconsciente y que no siempre son verbales [1]. Las expresiones faciales, la postura, los gestos, el tono de voz son algunos signos que nos permiten diferenciar las emociones y condicionan la interacción con el medio.

En palabras de Rosalind Picard [5]:

> *"Si queremos que los computadores sean realmente inteligentes y que interactúen de forma natural con nosotros, debemos otorgarles la capacidad de reconocer, entender e incluso tener y expresar emociones."*

Queda clara la importancia de las emociones en la vida humada, pero debemos estudiar su significado para comprender su esencia e interpretar sus objetivos al interactuar con el medio. La principal incógnita que se nos presenta es si se puede crear un sistema capaz de reconocer las emociones a través de la voz. Esta es la motivación principal de este proyecto y el fin es mejorar la interacción hombre-máquina, minimizando un poco más la brecha entre ambos.

1.1.1. Significado de la palabra emoción

Aunque las emociones están presentes cotidianamente en todo aspecto de la vida humana, tratar de definirlas es un trabajo costoso, que difiere según el autor y el contexto [6, 7, 8]. A pesar de ser estudio de múltiples disciplinas, no se ha logrado llegar a una definición unívoca.

Etimológicamente, la palabra emoción proviene del latín *emovere* que significa "salir hacia afuera". La Real Academia Española [9], define emoción como "alteración del ánimo intensa y pasajera, agradable o penosa, que va acompañada de cierta conmoción somática".

En muchos casos se describen las emociones utilizando sinónimos tales como sentimiento, afecto, motivación o estado de ánimo [10]. Desde Darwin [6] hasta la actualidad, el estudio de las emociones ha crecido de forma exponencial, originando así diversas perspectivas sobre el significado de la palabra emoción. Algunos la definen como cambios fisiológicos y neurológicos a causa de estímulos externos; desde la perspectiva biológica, definen

emoción como cambios en las expresiones faciales; otros la definen como la interacción de cambios fisiológicos y la evaluación cognitiva; otra rama la enfoca sólo desde la evaluación cognitiva de estímulos u objetos.

1.1.2. La voz de las emociones

Las emociones nos informan de lo que nos pasa en un momento determinado, brindándonos un estado afectivo. Ante cada situación, nos aportan información relevante para evaluar y actuar en consecuencia. Existen diferentes emociones que surgen según el momento que atraviesa la persona, que aportan un mensaje y una intensidad especial. El ser humano experimenta que cada una de las emociones emerge de su interior con el fin de comunicar algo. Así por ejemplo la *tristeza* nos informa de una meta perdida o no alcanzada, la *felicidad* nos expresa que hemos logrado alcanzar con éxito nuestras metas, la *ira* nos informa que estamos en presencia de una situación injusta, la *sorpresa* nos comunica que estamos frente a un acontecimiento inesperado, el *asco* nos advierte de la necesidad de evitar o alejarnos de un objeto o estímulo, el *miedo* nos anuncia la falta de seguridad [11].

En muchas ocasiones los individuos tienden a disimular o evitar expresar una emoción. La realidad es que detrás de la voz de cada una de las emociones, se tiende a expresar una necesidad, una realidad que resulta esencial para convivir y relacionarse con los demás.

Podrían listarse algunas peculiaridades importantes de las emociones:

- Manifiestan el valor de una situación para el individuo.
- Disponen al individuo a responder ante una situación.
- Advierten el estado y las intenciones.
- Ayudan a una mejor comunicación entre locutores.

Algunas investigaciones [12] afirman que alrededor del 90 % de sus manifestaciones son a través de la expresión no verbal, como ser en los gestos, la postura, el tono de voz, las expresiones faciales, la temperatura corporal, etc. Esta tendencia en la mayoría de los casos es captada implícitamente por el interlocutor, asimilando y respondiendo del mismo modo [3].

1.1.3. Papel de las emociones en la interacción hombre - computadora

Como se señaló en la sección anterior, el ser humano experimenta emociones que están presentes en todos los órdenes de su vida, aunque en la

mayor parte del tiempo se manifiestan de forma inconsciente. Al momento de interactuar un individuo con las computadoras, las emociones también están presentes. Es en este momento cuando los humanos notan con mayor atención las emociones que están experimentando, ya que se enfrentan a dispositivos que no tienen la capacidad de interactuar del mismo modo que un interlocutor humano. Como en toda actividad humana están presentes las emociones, también se deberían tener en cuenta en el momento de la interacción con máquinas [13].

Los seres humanos transmiten información emocional de manera intencional y no intencional a través de la voz. Estos patrones vocales son percibidos y comprendidos por los oyentes durante la conversación, pero no se experimenta lo mismo al interactuar con una computadora. Los seres humanos sienten que la comunicación con otros seres humanos es más natural, a causa de que la información adicional representada en sus expresiones emocionales, pueden ser reconocidas, tratadas y reflejadas. Por lo tanto, cuando los usuarios interactúan con los sistemas informáticos, existe una diferencia entre la información transmitida y la información percibida por la máquina.

A causa del desarrollo tecnológico actual, las interfaces están cada vez más cerca de las personas y por esto, el papel de las emociones se debe considerar. Como la interacción es cada día más frecuente y el abanico de usuarios es muy amplio, se requieren más y mejores máquinas con características afectivas que mejoren la interacción con los humanos.

1.2. Antecedentes

Gran parte de los trabajos recientes en este campo han centrado su análisis en las características prosódicas del discurso [14] y en las características espectrales de la voz [15, 16]. Entre los métodos más investigados y empleados para el reconocimiento de emociones están las máquinas de soporte vectorial (SVM)[17] y los modelos de mezclas de gaussianas (GMM, del inglés *Gaussian Mixture Models*)[18]. Sin embargo, también se han explorado aunque en menor medida, la utilización de modelos ocultos de Markov (HMM, del inglés *Hidden Markov Models*)[15]. En [15] se propone el uso de un vector de 39 características para modelar las señales de voz, entre éstas están la energía, la frecuencia fundamental y las formantes. Se emplean dos métodos de clasificación, máquinas de soporte vectorial y modelos ocultos de Markov, con los que se clasifican cinco estados emocionales. Con SVM se obtuvo un porcentaje de aciertos del 93.6 % sobre frases de hablantes femeninos y del 89.4 % sobre frases de hablantes masculinos, mientras que con los

HMM se obtuvo un reconocimiento del 98.9 % y del 100 % para hablantes femeninos y masculinos respectivamente. El trabajo [19] basa su estudio en el análisis de dos características: la frecuencia fundamental y la energía, y utiliza modelos ocultos semicontínuos de Markov. Con la combinación óptima de características de bajo nivel y de estructuras de HMM, han alcanzado un 80 % de aciertos, involucrando siete emociones. El corpus de emociones fue realizado por dos actores profesionales, un actor y una actriz. Otro trabajo de reconocimiento de emociones interesante es [16], donde se han construido tres clasificadores diferentes, variando las características modeladas y los métodos de clasificación. En un primer modelo se utilizan características espectrales (coeficientes Mel-Cepstrum, primera y segunda derivadas) y un GMM de 512 componentes gaussianas, logrando una tasa de aciertos del 98.4 %. En el segundo, se utilizan sólo 6 parámetros prosódicos y SVM, se obtiene una precisión del 92.32 %. El último modelo se vale de 6 características prosódicas y GMM para obtener 86.71 % de aciertos. Estos resultados fueron obtenidos con un corpus de seis emociones grabado por una actriz, lo cual supone una dependencia del hablante.

Clasificación de las emociones

Al igual que la definición de *emoción* tiene múltiples variantes, su clasificación también fue evaluada por diferentes autores sin arribar a una clasificación unívoca, como ya se comentó previamente. Actualmente, se pueden encontrar en la literatura diversas clasificaciones basadas en diferentes criterios.

Desde el punto de vista psicológico, los autores clasifican las emociones como agradables/desagradables o emociones positivas/negativas o emociones con alto nivel de activación/con bajo nivel de activación [20, 21]. Es decir, cada emoción presenta una medición que puede tomar valores positivos o negativos. Ejemplo de emociones negativas son: la tristeza, el enojo, la vergüenza, el miedo; y de emociones positivas son: la alegría, la esperanza, el amor.

Otra rama de la psicología clasifica las emociones en primarias y secundarias. Las primeras son emociones fundamentales o básicas mientras que las secundarias pueden ser vistas como derivadas a partir de las primarias. Es posible distinguir, en la categoría primaria, a las *seis grandes* emociones [7, 22] : alegría, sorpresa, miedo, asco, enojo y tristeza. Entre las emociones secundarias se pueden encontrar: burla, melancolía, euforia, sarcasmo, entre otras. Esta clasificación es análoga a la forma de clasificar a los colores, don-

de se los divide en primarios y secundarios, y la mezcla de colores primarios forma un color secundario. Dependiente de la cantidad de los colores que se mezclen, será el color que se obtendrá como resultante. Lo mismo sucede con las emociones secundarias, suelen aparecer como una combinación de las emociones primarias, que dependen de la intensidad de las emociones primarias que las componen.

A continuación se repasan algunas de las características principales de las seis emociones primarias y de algunas emociones secundarias [8].

Emociones primarias

- Enojo: no hay un consenso general sobre los efectos que se generan en esta emoción. El enojo se caracteriza por aumentar la actividad fisiológica y neurológica. Desde el punto de vista de las características prosódicas presenta un tono medio alto, un amplio rango de tono y la velocidad de la locución es rápida.

- Alegría: generalmente esta emoción se experimenta por un lapso corto de tiempo y surge al considerar que un estímulo u objeto son favorables para el logro de una meta deseable para el individuo. Las características prosódicas que se registran en esta emoción es un tono alto, mayor velocidad de locución y la entonación se registra ascendente y descendente a lo largo de la frase, produciendo un elevamiento al final.

- Tristeza: esta emoción está ligada a la pérdida irremediable de una meta u objeto que produce un decaimiento en el estado de ánimo del individuo. En el ámbito de las características prosódicas, se caracteriza por presentar un tono bajo, velocidad de locución lenta y la entonación final descendiente.

- Miedo: se experimenta en los momentos en que el individuo se da cuenta que su vida se encuentra amenazada. Es el deseo de evitar que algo suceda, produciendo un aumento del ritmo cardíaco. Al observar las características prosódicas, se manifiesta un tono elevado sobre la normal con variaciones ascendentes y descendentes, rango elevado y rápida velocidad de elocución.

- Asco: es la emoción ligada al gusto, en el sentido de repugnancia a un objeto o sustancia. Se caracteriza por presentar un tono medio bajo, amplio rango y velocidad de locución lenta con pausas extensas.

- Sorpresa: es un estado emocional breve y a menudo involuntario, que es el resultado de experimentar un evento inesperado. Presenta cambios en las expresiones faciales y, prosodicamente, el tono es mayor al normal, rango amplio y velocidad igual a la normal.

Emociones secundarias

Entre las emociones secundarias, se encuentran: pena, ternura, ironía, culpa, vergüenza, orgullo, celos, queja, anhelo, aburrimiento, satisfacción, impaciencia, etc. Debido a que el número de emociones que caen en esta clasificación es muy alto, sólo unas pocas se caracterizan a continuación.

- Pena: es el sufrimiento, mental o físico a causa de una lesión o pérdida. Sus características prosódicas se presentan con un tono bajo, rango estrecho y una lenta velocidad de locución con grandes cantidades de pausas.
- Ternura: se manifiesta en el individuo al vivir una situación extrema de paz y armonía con otra persona. Desde el punto de vista prosódico, su tono se presenta alto pero sin grandes variaciones.
- Ironía: es expresada por el locutor con la intención de comunicar lo contrario de su significado. Se caracteriza por presentar una velocidad de locución baja.

1.3. Aplicaciones del reconocedor de emociones

Son múltiples los usos que se pueden dar al reconocedor y los ámbitos a los que se puede integrar. Un ejemplo son los centros de llamadas. En muchas ocasiones se acude con la necesidad de determinado servicio y es una máquina la que responde la llamada. Hoy en día, muchas personas se rehúsan a interactuar con máquinas. Esto puede dar lugar a un descontento por parte del cliente y como el estado emocional del locutor no es percibido por la máquina puede producir, en el peor de los casos, la pérdida de un cliente. Si a la interacción entre el cliente y la máquina se adiciona el reconocimiento de emociones y se detecta en una llamada que el locutor está enojado o confundido con el sistema automatizado, su llamada se puede redireccionar a un operador humano para obtener asistencia.

Otro ámbito de aplicación puede ser en centros de emergencia dispuestos en estaciones de policía, estaciones de bomberos, centros de asistencia médica, etc. Cada día, muchos de estos centros reciben llamadas de broma y

es el telefonista quien debe verificar si la persona está en una emergencia o no. Sin ir mas lejos, en el centro de llamadas del 911 de nuestra ciudad se han registrado un 74 % de llamadas falsas, en un período de tres meses[1]. Si en estos ámbitos se integrara el reconocedor de emociones, se podría determinar de forma objetiva si la persona que llama presenta un estado emocional de emergencia (susto, enojo, tristeza, etc.). Esto permitiría tomar decisiones mucho más eficaces y rápidas en la atención de la comunidad.

La incorporación del reconocedor podría traer grandes ventajas en las sesiones terapéuticas debido a que determinadas patologías se asocian a cambios emocionales. Muchos pacientes, al momento de la sesión, no demuestran realmente la emoción que poseen interiormente. Por lo tanto, es muy importante la percepción y experiencia del terapeuta/psicólogo para saber realmente el estado emocional de la persona. Si el paciente es sometido a una prueba con el reconocedor al momento de la sesión, se podría pre-diagnosticar una patología de forma más eficaz y precoz ya que el terapeuta sabría el estado emocional verdadero de la persona. Otro ejemplo concreto de aplicación sería en el estudio de niños autistas.

En el caso de las "casas inteligentes", el uso de un reconocedor de emociones serviría para el bienestar de las personas que viven en el hogar. Por ejemplo, si uno de los habitantes está triste, se dispararía un evento que pondría música alegre. Si en cambio estuviera enojado, se reproduciría una melodía más relajante. El reconocedor de emociones en este caso aumentaría la confortabilidad del usuario.

1.4. Objetivos del proyecto

El objetivo general puede definirse como

"Desarrollar una herramienta capaz de analizar las señales de voz emitidas por el ser humano, para reconocer emociones de forma automática, con el fin de lograr una interacción hombre-máquina más natural."

Objetivos específicos

- Desarrollar rutinas de procesamiento de señales para el análisis de características espectrales. Evaluar las señales de voz con el fin de hallar información implícita que sirva para caracterizar las emociones.

[1]Nota publicada en el diario El Litoral de la ciudad de Santa Fe el día 11 de julio de 2009

- Registrar una base de datos de emociones en español.
- Aplicar tecnologías de procesamiento de señales e inteligencia computacional a un problema práctico de interés en la ingeniería informática.
- Desarrollar dos sistemas de reconocimiento, en base a mezclas de gaussianas y modelos ocultos de Markov. Comparar el desempeño de ambos para el reconocimiento de hasta 7 emociones expresadas en la voz.
- Realizar una completa y detallada documentación de cada etapa con sus respectivas tareas.

Capítulo 2

Extracción de características y métodos de clasificación

En este capítulo se hace una descripción detallada de las principales técnicas de extracción de características y métodos de clasificación que se utilizaron en el presente proyecto. El objetivo principal es proponer un marco teórico como fundamento del desarrollo que se presenta en los capítulos posteriores. Este capítulo se divide en tres grandes bloques especialmente orientados y restringidos a su utilización en el reconocimiento automático de emociones: el análisis de la señal de voz, la mezcla de gaussianas y los modelos ocultos de Markov.

En primer lugar se trata, como marco general, el análisis por tramos de la señal de voz. A partir de esta particular forma de seguir la dinámica de la voz, se describen los diferentes métodos de análisis. En la segunda parte del capítulo se describe el modelo de mezcla de gaussianas, y en la tercera se describe la estructura de un sistema de reconocimiento automático de emociones basado en modelos ocultos de Markov.

2.1. Análisis de señales de voz

Dentro de las características morfológicas de la señal de voz están la de ser variable en el tiempo y contínua. Es decir, por más pequeño que sea el intervalo que se analice, siempre se obtendrá un valor de la señal [23].

Como la mayoría de los sistemas que procesan señales de voz utilizan una computadora, se debe convertir la señal contínua en una señal digital a través de un proceso conocido como digitalización de la señal [23]. Esta conversión la llevan a cabo sistemas de conversión analógica a digital, que

toman una muestra cada cierto intervalo de tiempo y realizan la cuantización de la amplitud de la señal en valores discretos. Este proceso también suele denominarse *muestreo* y *cuantización* de la señal de voz.

En estas señales de voz digitalizadas, se hace valer la hipótesis de estacionariedad por tramos. Su validez se explica en relación a la velocidad de variación de la morfología del tracto vocal, entonces se considera a la señal de voz estacionaria en períodos de 10 a 30 ms aproximadamente [24]. Para realizar el análisis por tramos, se utilizan diversos tipos de ventanas de análisis, entre ellas se pueden citar: la ventana Hanning, la ventana de Hamming y la ventana de Blackman [25]. Estas ventanas ayudan a disminuir el fenómeno de distorsión armónica, conocido como el fenómeno de Gibbs, que genera una ventana rectangular en el proceso de ventaneo.

Las ventanas pueden ser caracterizadas por el tamaño de los lóbulos de la magnitud de su espectro de frecuencias. Por ejemplo ventana rectangular posee un lóbulo central con un ancho de banda pequeño pero la magnitud de los lóbulos laterales decae muy lentamente, mientras que una ventana de Blackman posee mínima amplitud en sus lóbulos laterales pero su lóbulo principal tiene un ancho de banda tres veces mayor al de la rectangular [25]. La elección del tipo de ventana, si bien está asociado al tipo de aplicación, debe basarse en un compromiso entre la capacidad de reducción del fenómeno de Gibbs y el costo computacional asociado a la utilización de la ventana. Para el análisis de señales de voz suele utilizarse la ventana de Hamming, pues ofrece el compromiso más adecuado [26].

Las señales de voz no brindan información importante en el dominio temporal, por ello es necesario transformarlas, para extraer características útiles al análisis [27]. Teniendo en mente que estas transformaciones se realizan a esos tramos ventaneados de la señal, a continuación se comentan brevemente algunas de las transformaciones más utilizadas, en su forma más general.

Coeficientes espectrales (CE): se aplica la transformada de Fourier (FT) a cada tramo de la señal y se obtienen las características frecuenciales de la misma:

$$ce(k) = \sum_{n=1}^{N_x} x(n) e^{-\mathrm{j}(2\pi/N_x)k(n-1)}$$

donde $x(n)$ es la señal en estudio.

Coeficientes ceptrales (CC): el análisis cepstral es un caso especial de los métodos de transformación homomórficos [28]. Se aplica en el análisis de señales de voz para extraer la información del tracto vocal. Basándose en la FT [23], el ceptrum se define como la aplicación de la FT inversa al logaritmo de la FT de la señal de voz en estudio:

$$cc(n) = FT^{-1}\{\log|FT\{x(n)\}|\}.$$

Coeficientes ceptrales en escala de mel (MFCC): un *mel* es una unidad de medida de la frecuencia fundamental percibida. Surge de la percepción de tonos puros en el ser humano. La ecuación que permite el mapeo entre las frecuencias en escala real (Hz) y en las frecuencias en escala perceptiva (mel)[26] es:

$$F_{mel} = 1000 \log_2 \left[1 + \frac{F_{Hz}}{1000}\right]$$

Para obtener los coeficientes MFCC, se calcula la FT de la señal de voz y el espectro es mapeado en la escala de mel. Luego, se aplica el logaritmo a cada tramo de las frecuencias en el dominio de mel y se calcula la FT inversa. En general, la *FT* inversa se sustituye por la transformada coseno (TC) con el fin de simplificar el cálculo.

Coeficientes de energía, delta y aceleración

Al conformar el vector de características, usualmente se incluyen coeficientes que dan más información sobre el tramo de la señal en estudio. Entre ellos se encuentra la *energía*, la cual se define como:

$$E = \log \sum_{n=1}^{N_x} x^2(n) \tag{2.1}$$

Los *coeficientes delta y aceleración* corresponden a una aproximación a la primera y segunda derivada temporal. Para un vector de características $x(k)$ dado, se obtienen los coeficientes delta mediante la regresión [29]:

$$\Delta x_t(k) = \frac{\sum_{j=1}^{N_J} j[x_{t+j}(k) - x_{t-j}(k)]}{2\sum_{j=1}^{N_J} j^2} \tag{2.2}$$

Los coeficientes de aceleración $\Delta^2 x_t(k)$ se obtienen por aplicación directa de la ecuación anterior a los $\Delta x_t(k)$.

2.2. Mezcla de gaussianas

Aunque las distribuciones gaussianas tienen propiedades analíticas importantes, se evidencian sus limitaciones al modelar datos reales. Si los datos reales que estamos analizando se concentran en dos grupos bien distanciados, una distribución gaussiana simple no capturará su estructura correctamente, mientras que la superposición de dos distribuciones se ajustaría mejor a la distribución real de los datos. Las superposiciones mencionadas, formadas como combinación lineal finita de distribuciones más simples, dan lugar a los llamados *modelos de distribuciones mezcladas* o *modelos de mezclas* que se utilizan ampliamente para la estimación en estadística. Si cada distribución simple es una densidad de gaussiana, el modelo se denomina *Mezcla de gaussianas* (GMM, del inglés *Gaussian Mixture Models*) [18] y tiene la forma:

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k), \tag{2.3}$$

donde $\mathcal{N}()$ es un función normal, el vector de media $\boldsymbol{\mu}_k$ y la matriz de covarianza Σ_k; K es el número de gaussianas y c_k los pesos asociados con cada una de las funciones, cumpliéndose que

$$\sum_k c_k = 1 \tag{2.4}$$

y $0 \leq c_k \leq 1$ para todo k.

Cada gaussiana $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$ en el modelo de mezclas de gaussianas permite obtener una aproximación burda de las características originales observadas. Mientras que la superposición de éstas permite una muy buena aproximación.

Si planteamos el modelo en base a sus parámetros como $\lambda = \{\boldsymbol{\mu}_k, \Sigma_k, c_k\}$ con $k = 1, 2, \ldots, K$, entonces éste queda determinado por el vector de medias $\boldsymbol{\mu}$, la matriz de covarianza Σ y el vector de pesos de la distribución $\mathbf{c}$. Estos parámetros pueden determinarse a través del algoritmo de maximización de la esperanza (EM) [18], donde se comienza con una estimación inicial de los parámetros $\lambda(0)$ a partir de la que se estiman nuevos parámetros del modelo $\lambda(1)$. Esto se hace iterativamente hasta alcanzar cierto criterio de convergencia.

Para una observación $\mathbf{x}_n$ se puede expresar (2.3) como

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} p(k)p(\mathbf{x}_n|k), \tag{2.5}$$

donde $p(k) = c_k$ y $p(\mathbf{x}_n|k)$ es la k-ésima distribución normal.

Luego, por el teorema de Bayes, la probabilidad a posteriori se puede escribir como

$$\gamma_{nk} \equiv p(k|\mathbf{x}_n) = \frac{p(k)p(\mathbf{x}_n|k)}{\sum_l p(l)p(\mathbf{x}_n|l)} = \frac{c_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_l c_l\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_l, \Sigma_l)} \tag{2.6}$$

Si se desea explicar la distribución de un conjunto de observaciones $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ mediante GMM. Este conjunto de datos se representan mediante una matriz $\mathbf{X}$ de dimensiones NxD y filas $\mathbf{x}_n'$. La función de verosimilitud logaritmica está dada por

$$\log p(\mathbf{X}|k) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} c_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right\} \tag{2.7}$$

Se maximiza la función - $\log p(\mathbf{X}|k)$ igualando a cero su derivada respecto de $\boldsymbol{\mu}_k$, Σ_k y c_k, para obtener las fórmulas de reestimación de los coeficientes del modelo

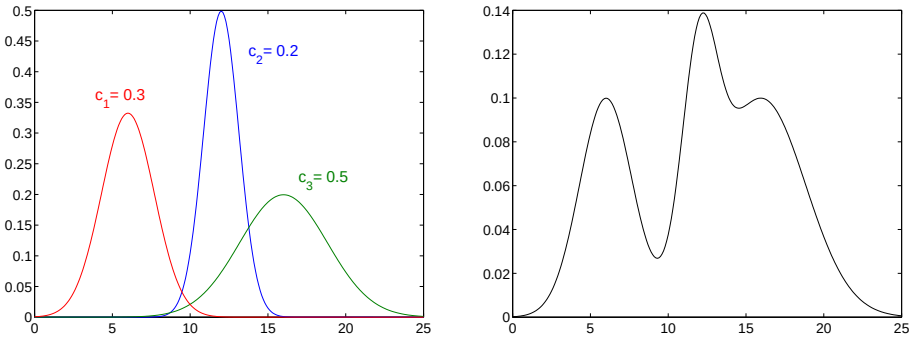

Figura 2.1. Distribución modelada por la mezcla de 3 gaussianas simples en una dimensión.

$$\tilde{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n \tag{2.8}$$

$$\tilde{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \tilde{\boldsymbol{\mu}}_k)(\mathbf{x}_n - \tilde{\boldsymbol{\mu}}_k)^T \tag{2.9}$$

$$\tilde{c}_k = \frac{N_k}{N} \tag{2.10}$$

con $N_k = \sum_{n=1}^{N} \gamma_{nk}$. Usando un número suficiente de gaussianas, ajustando sus medias, covarianzas y los coeficientes de la combinación lineal, se puede aproximar casi cualquier tipo de densidad continua con una precisión arbitraria [18].

En la Figura 2.1 (extraídas de [18]) se puede observar una mezcla gaussianas unidimensional formada por tres gaussianas simples, cada una escalada por un coeficiente c_k.

Los modelos de mezclas de gaussianas fueron aplicados en el reconocimiento automático del habla y su uso fue extendido al reconocimiento de emociones ya que permiten obtener buenos resultados [16].

2.3. Modelos ocultos de Markov

Los modelos ocultos de Markov (HMM, del inglés *Hidden Markov Models*) constituyen una técnica de modelado estocástico que ha cobrado gran interés por su favorable desempeño en el procesamiento de señales, más

particularmente en el área de reconocimiento del habla (ASR, del inglés *Automatic Speech Recognition*) [30]. En los años 60 se empezaron a desarrollar estos modelos pero su mayor auge fue durante los últimos años en su uso en problemas de clasificación de señales.

Los modelos ocultos de Markov tratan de caracterizar alguna propiedad estadística de la señal. En estos modelos se asume que la señal se puede caracterizar perfectamente como un proceso aleatorio paramétrico y que los parámetros se pueden calcular o estimar.

Un HMM está compuesto de dos elementos básicos: un proceso de Markov y un conjunto de distribuciones de probabilidad de salida. El modelo puede ser considerado una máquina de estados finitos probabilística, ya que contiene un conjunto de estados conectados unos a otros por arcos de transición y probabilidades asociadas a cada arco. Así, los HMM pueden utilizarse para modelar cualquier serie o secuencia temporal.

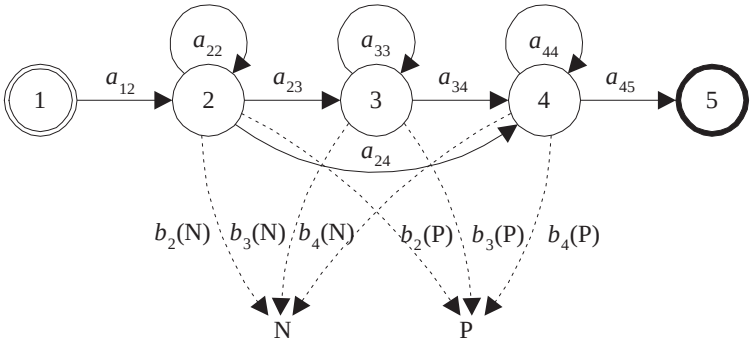

Figura 2.2. Diagrama de un modelo oculto de Markov de 5 estados.

En la Figura 2.2 [31] se puede observar un HMM con una red de estados interconectados comúnmente utilizado en RAH, los cuales emiten de acuerdo a ciertas funciones de probabilidad. En cualquier instante de tiempo especificado se puede considerar que el sistema está en uno de los estados disponibles y que, a intervalos regulares de tiempo, ocurre una transición a otro estado (o al mismo estado si este dispone de una transición a si mismo)[28]. Esto se produce conforme a las probabilidades asociadas a los arcos de transición. Los HMM más utilizados en ASR poseen una estructura muy simple denominada de izquierda a derecha, donde las transiciones se dan solamente en ese sentido.

El parámetro $b_j(k)$ especifica que la probabilidad de que el estado j

observe el símbolo k del conjunto de símbolos observables. Como se ve en la figura, todos los estados pueden emitir todos los símbolos, en estas condiciones nunca se podrá saber con certeza en que estado esta el modelo observando solamente su salida. Por lo tanto, el funcionamiento interno del modelo queda "oculto" y es por eso que se lo denomina modelo *oculto* de Markov.

Un HMM contínuo (CHMM) es definido mediante una estructura algebraica [30]:

$$\Theta = \langle \mathcal{Q}, \mathbb{O}, \mathbf{A}, \mathcal{B} \rangle \tag{2.11}$$

donde:

$\mathcal{Q}$ es el conjunto de estados posibles,

$\mathbb{O}$ es el espacio observable, los símbolos observados se corresponden con la salida física del sistema que se desea modelar,

$\mathbf{A}$ es la matriz de probabilidades de transición de estados y

$\mathcal{B}$ es el conjunto de distribuciones de probabilidades de observación de un símbolo en cada estado.

En el caso de los CHMM, en lugar de tener distribuciones de probabilidad discretas $b_j(i)$, para cada estado i, se modela una distribución de probabilidades expresadas mediante una combinación lineal de distribuciones gaussianas [29]:

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} c_{jk} b_{jk}(\mathbf{x}) \quad \forall j \in \mathcal{Q};\ K < \infty \tag{2.12}$$

donde

I) $N_x \in \mathbb{N}$ es el espacio de las evidencias acústicas,

II) K es el número de componentes de la mezcla,

III) $c_{jk} \in \mathbb{R}^{+0}$ es el coeficiente para la k-esima mezcla en el estado j, que satisface:

$$\sum_{k=1}^{K} c_{jk} \stackrel{\circ}{=} 1 \quad \forall j \in \mathcal{Q}, \tag{2.13}$$

IV) b_{jk} todas funciones de distribuciones normales, con la forma:

$$\mathcal{N}\left(x_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}\right) = \frac{1}{(2\pi)^{N_x} \left|\mathbf{U}_{jk}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(x_t - \boldsymbol{\mu}_{jk})^{\mathrm{T}} \mathbf{U}_{jk}^{-1} (x_t - \boldsymbol{\mu}_{jk})\right]}, \tag{2.14}$$

con

$\boldsymbol{\mu}_{jk} \in \mathbb{R}^{N_x}$: los vectores de medias,
$\mathbf{U}_{jk} \in \mathbb{R}^{N_x \times N_x}$: las matrices de covarianza

V) se cumple que:

$$\int_{-\infty}^{+\infty} b_j\left(x_t\right) \mathrm{d}x_t \doteq 1 \quad \forall j \in \mathcal{Q}. \tag{2.15}$$

En los modelos ocultos de Markov existen tres problemas a resolver para que resulten modelos útiles en aplicaciones reales [30]:

- *Problema de evaluación*: a partir de una secuencia de observaciones y un modelo, se busca calcular la probabilidad de que la secuencia haya sido producida por el modelo definido. La principal preocupación en este tipo de problemas es la eficiencia computacional.
- *Problema de estimación*: a partir de una secuencia de observaciones y un modelo, se busca como elegir una secuencia de estados que sea óptima en algún sentido, es decir, que mejor explique las observaciones. No existe una solución única para este problema ya que es dependendiente del criterio utilizado. Entre ellos, en este caso se utilizó la secuencia de estados más probables.
- *Problema de entrenamiento*: a partir de una o varias secuencias observadas, se busca cómo ajustar los parámetros del modelo en forma óptima.

En el siguiente apartado, se describirán las soluciones a estos tres problemas para el caso simplificado de los modelos discretos.

2.3.1. Problema de evaluación

Dado un modelo Θ, se desea calcular la probabilidad $p\left(\mathbf{X}^T|\Theta\right)$ de una secuencia de observables

$$\mathbf{X}^T = x_1, x_2, \ldots, x_T \tag{2.16}$$

donde cada x_t es un símbolo observado. Para eso, se enumeran todas las secuencias de estados posibles de longitud T. Luego, para una secuencia de estados dada

$$\mathbf{q}^T = q_1, q_2, \ldots, q_T; \quad q_t \in \mathcal{Q} \tag{2.17}$$

la probabilidad de la secuencia observada $\mathbf{X}^T$, dada la secuencia de estados y el modelo, es

$$p\left(\mathbf{X}^T\left|\mathbf{q}^T, \Theta\right.\right) = \prod_{t=1}^{T} p\left(x_t\left|q_t, \Theta\right.\right) \tag{2.18}$$

Por otro lado, la probabilidad correspondiente a la secuencia de estados $\mathbf{q}^T$ puede escribirse como

$$p\left(\mathbf{q}^T\left|\Theta\right.\right) = a_{q1q2}a_{q2q3}\ldots a_{q_{T-1}}a_{q_T} \tag{2.19}$$

La probabilidad de tal secuencia de observaciones y estados, es simplemente el producto de ambos

$$p\left(\mathbf{X}^T, \mathbf{q}^T\left|\Theta\right.\right) = p\left(\mathbf{X}^T\left|\mathbf{q}^T, \Theta\right.\right) p\left(\mathbf{q}^T\left|\Theta\right.\right) \tag{2.20}$$

Así, dado el modelo, la probabilidad de que la secuencia observada haya sido producida por dicho modelo, se puede obtener sumando la probabilidad conjunta sobre todas la secuencias de estados posibles

$$p\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T} p\left(\mathbf{X}^T\left|\mathbf{q}^T,\Theta\right.\right) p\left(\mathbf{q}^T|\Theta\right) \tag{2.21}$$

El orden de cálculo de ésta probabilidad es $2TN^T$, lo cual lo hace inaceptable computacionalmente. Esto se debe a que N^T son la cantidad de posibles secuencias de estados y tenemos $2T$ cálculos para cada secuencia. Para solucionar este inconveniente se utiliza el algoritmo Adelante-Atrás (Forward-Backward [32]). Vale destacar que sólo es necesario el apartado *hacia adelante* para dar solución al problema de evaluación, pero se explicará también el apartado *hacia atrás* ya que es utilizado en la solución al problema de entrenamiento.

Evaluación hacia adelante

Se define la variable hacia adelante como

$$\alpha_t(i) = p\left(x_1 x_2 \dots x_t, q_{t=i}|\Theta\right) \tag{2.22}$$

es decir, la probabilidad de generar la secuencia parcial de observaciones hasta que el modelo se encuentre en el estado i en el instante t. De esta manera se obtiene la probabilidad de que la secuencia parcial observada haya sido producida por el modelo HMM, pasando por el estado i en el tiempo t.

El algoritmo hacia adelante se puede resumir en los siguientes pasos:

1) Inicialización: para cada uno de los estados se inicializa $\alpha_1(i)$, en el instante $t = 1$, como la probabilidad conjunta del estado i y la observación inicial x_1. Para el primer estado siempre se inicia en 1 dado que la estructura del modelo es de izquierda a derecha.

$$\alpha_1(i) = b_i(x_1) \quad i = 1 \dots N \tag{2.23}$$

2) Inducción: el número de caminos necesarios para calcular α crece exponencialmente a medida que crece el número de secuencia de observaciones parciales. Pero α en el instante $t-1$ da la probabilidad de llegar al estado a través de todas las rutas anteriores, por lo tanto se puede definir la probabilidad α en el instante de tiempo t en términos de la anterior, en

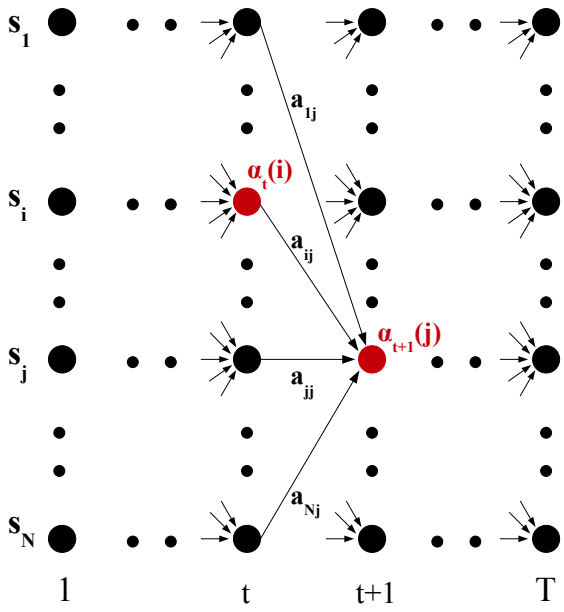

Figura 2.3. Cálculo de la variable hacia adelante.

el instante $t-1$. Se calcula $\alpha_{t+1}(j)$ como el producto de la probabilidad de observación de x_{t+1} en el estado j (es decir, $b_j(x_{t+1})$) por la suma de las probabilidades de secuencia parcial en el instante t, multiplicadas por las probabilidades de transición entre este estado y j (es decir, a_{ij}). Este proceso puede verse esquematizado en la Figura 2.3.

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} [\alpha_t(i) a_{ij}] b_j(x_{t+1}) \quad t = 1, \ldots, T-1; \; j = 1, \ldots, N \qquad (2.24)$$

3) Terminación. La última observación puede producirse en cualquiera de los estados. El algoritmo culmina sumando todas las variables *hacia adelante* $\alpha_T(i)$ en el instante final T, obteniendo la probabilidad total

$$p\left(\mathbf{X}^T|\Theta\right) = \sum_{i=1}^{N} \alpha_T(i) \tag{2.25}$$

En la Figura 2.3 se aprecian los estados y las probabilidades necesarias para el cálculo $\alpha_t(i)$. El cálculo de probabilidad mediante éste algoritmo tiene un orden N^2T, lo cual lo hace más eficiente frente a la evaluación exhaustiva, de orden $2TN^T$.

Evaluación hacia atrás

Consideramos la variable hacia atrás como

$$\beta_t(i) = p\left(x_{t+1}, x_{t+2}, \ldots x_T | q_{t=i}, \Theta\right) \tag{2.26}$$

es decir, dado el estado i en el instante t y el modelo, la probabilidad de secuencia observada desde el instante $t+1$ hasta el final. Por inducción, el cálculo de la variable hacia atrás se calcula como sigue:

1) Inicialización: se define la variable hacia atrás

$$\beta_T(i) = 1 \quad i = 1, \ldots, N \tag{2.27}$$

2) Inducción: de manera análoga al cálculo por inducción hacia adelante se calculan los $\beta_t(i)$. Se debe considerar que del estado i se puede ir a cualquiera de los otros estados definidos en el modelo.

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, \ldots 1;\ i = 1, \ldots N \tag{2.28}$$

3) Terminación: la primera observación se produce en el estado 1 dado que el modelo es de izquierda a derecha. La probabilidad buscada se obtiene

$$p\left(\mathbf{X}^T|\Theta\right) = b_1(x_1)\beta_1(1) \tag{2.29}$$

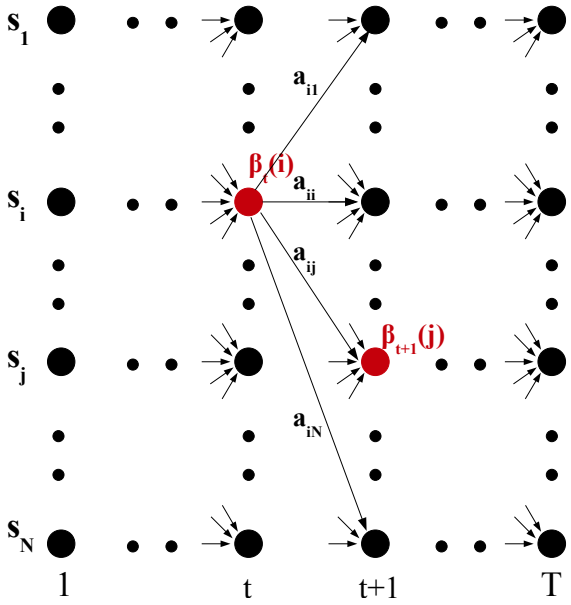

Figura 2.4. Cálculo de la variable hacia atrás.

En la Figura 2.4 se aprecian los estados y las probabilidades necesarias para el cálculo $\beta_T(i)$. Esta evaluación requiere un costo computacional de orden N^2T. De este modo, tanto la evaluación hacia adelante como la evaluación hacia atrás, tienen el mismo orden de complejidad, que es muy inferior al del cálculo directo.

2.3.2. Problema de estimación

No existe una solución única para tomar la secuencia de estados que sea óptima en algún sentido. Como se hizo referencia anteriormente, depende del criterio con que se defina esta secuencia óptima. Para la solución de este problema, se elegirá el criterio de selección de la secuencia de estados donde es máxima la probabilidad de observar secuencia de símbolos dada:

$$\tilde{\mathbf{q}}^T = \arg \max_{\forall \mathbf{q}^T} \left\{ p\left(\mathbf{X}^T \left| \mathbf{q}^T, \Theta \right.\right) \right\} = \arg \max_{\forall \mathbf{q}^T} \left\{ p\left(\mathbf{X}^T \, , \mathbf{q}^T | \Theta \right.\right) \tag{2.30}$$

Al igual que el cálculo directo para el problema de evaluación, el cálculo de (2.30) tampoco puede ser aplicada dada su complejidad computacional. Para subsanarlo, se utiliza un algoritmo recursivo análogo al de adelante-atrás, denominado algoritmo de Viterbi.

Algoritmo de Viterbi

El algoritmo de Viterbi se basa en técnicas de programación dinámica. A partir de la secuencia de observaciones $\mathbf{X}^T$, para encontrar la secuencia óptima de estados, la variable $\delta_t(i)$ se define como:

$$\delta_t(i) = \max_{\mathbf{q}_T} \left\{ p\left(q_1, q_2, \ldots, q_{t=i} \, , x_1, x_2, \ldots, x_t | \Theta\right) \right\} \tag{2.31}$$

La formulación del algoritmo de Viterbi es similar al algoritmo hacia adelante con la diferencia de la función máx en lugar de la sumatoria sobre todas las secuencias de estados posibles. Los pasos a seguir en el algoritmo, se describen a continuación:

1) Inicialización: se inicializa el algoritmo en el estado $j = 1$ y la observación x_1, es decir,

$$\delta_1(i) = b_i(x_1) \quad i = 1 \dots N \tag{2.32}$$

2) Inducción: este es el paso más importante, donde se obtiene la máxima probabilidad acumulada. Para cada uno de los estados e instantes de tiempo, se calcula la probabilidad del mejor camino $\delta_{t+1}(i)$

$$\delta_{t+1}(i) = \max_{1<i<N} \{\delta_t(i)a_{ij}\} b_j(x_{t+1}) \tag{2.33}$$

Como el objetivo consiste en obtener la secuencia de estados de probabilidad máxima, se almacenan los valores de (2.33) para cada t y j. Para eso se utiliza la matriz $\psi_t(j)$

$$\psi_t(j) = \arg \max_{1<i<N} (\delta_{t-1}(i)a_{ij}) \tag{2.34}$$

3) Terminación. Una vez obtenidos los valores para cada t y j, se procede a encontrar la secuencia de estados $\tilde{\mathbf{q}}_T$ asociada a la máxima probabilidad:

$$\tilde{q}_T = \arg \max_{1<i<N} \{\delta_T(i)\}$$

por recursión inversa:

$$\tilde{q}_t = \psi_{t+1}(\tilde{q}_{t+1}); \quad t = T-1, T-2, \dots, 1 \tag{2.35}$$

El algoritmo es análogo al *hacia adelante*, y aunque los valores obtenidos son diferentes, se suele utilizar el algoritmo de Viterbi para determinar la probabilidad de una secuencia de observaciones por el camino óptimo, ya que requiere menos operaciones.

2.3.3. Problema de entrenamiento

El tercer problema a resolver es el ajuste de los parámetros del HMM, para maximizar la probabilidad de que la observación haya sido generada por el modelo. La solución a éste problema se puede obtener por máxima verosimilitud (ML, del inglés *Maximum Likelihood*). Como el funcionamiento interno del modelo HMM está oculto, se calcula la estimación de ML a través del *algoritmo de Baum-Welch*, el cual es un caso particular del algoritmo de *maximización de la esperanza* (EM) [33]. El algoritmo de Baum-Welch permite estimar los parámetros de un modelo que hacen máxima la probabilidad de una secuencia de observaciones.

Reestimación de Baum-Welch

Se define la variable $\xi_t(i,j)$ como la probabilidad de que el modelo se encuentre en el estado i en el tiempo t y en el estado j en el tiempo $t+1$, dado el modelo Θ y la secuencia de observaciones $\mathbf{X}^T$

$$\xi_t(i,j) = p(q_t = i, q_{t+1} = j|\mathbf{X}^T, \Theta) = \frac{p(q_t = i, q_{t+1} = j|\mathbf{X}^T, \Theta)}{p(\mathbf{X}^T|\Theta)}, \quad (2.36)$$

En función de las variables *hacia adelante* y *hacia atrás* definidas previamente, esta probabilidad se puede expresar como [30]

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(x_{t+1})\beta_{t+1}(j)}{p(\mathbf{X}^T|\Theta)}, \quad (2.37)$$

Otra variable utilizada para el procedimiento de reestimación de Baum-Welch es $\gamma_t(i)$, que define como la probabilidad de estar en el estado i en el instante t, habiendo observado la secuencia completa $\mathbf{X}^T$

$$\gamma_t(i) = p\left(q_t = S_i \left|\mathbf{X}^T, \Theta\right.\right) = \frac{p(q_t = S_i|\mathbf{X}^T, \Theta)}{p(\mathbf{X}^T|\Theta)}, \quad (2.38)$$

Del mismo modo que $\xi_t(i,j)$, se expresa $\gamma_t(i)$ en función de las variables *hacia adelante y hacia atrás*:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{p(\mathbf{X}^T|\Theta)}, \tag{2.39}$$

Interpretando las fórmulas anteriores, si sumamos en cada instante $\gamma_t(i)$, nos da por resultado una estimación del número de veces que el modelo se encontró en el estado j para la secuencia $\mathbf{X}^T$ dada. Sumando la variable $\xi_t(i,j)$ hasta el instante $T-1$ se obtiene el número de veces que se produjo una transición entre los estados i y j, al observar la secuencia completa $\mathbf{X}^T$.

Basándose en estas expresiones se pueden obtener las fórmulas de estimación de los parámetros de un HMM.

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{2.40}$$

$$\tilde{c}_j(k) = \frac{\sum_{t=1}^{T} \xi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2.41}$$

$$\tilde{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) x_t}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{2.42}$$

$$\tilde{\mathbf{U}}_{jk}^{-1} = \frac{\sum_{t=1}^{T} \xi_t(j,k)(x_t - \tilde{\boldsymbol{\mu}}_{jk})(x_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}}}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{2.43}$$

Donde (2.40) es la fórmula de reestimación para las probabilidades de transición de estados; (2.41) es la constante de pesos relativos a cada distribución de probabilidad de observación, junto con la fórmula (2.42) que permite reestimar los vectores de medias y (2.43) las matrices de covarianza. A partir de un modelo Θ, utilizando las fórmulas de reestimación, se

genera un nuevo modelo $\tilde{\Theta}$. En el mismo se verifica que la probabilidad de generación de la secuencia de observaciones es superior a la inicial, salvo al alcanzar un valor crítico de la función de verosimilitud [32].

2.3.4. Modelo de lenguaje

El modelo de lenguaje es empleado para determinar que evento (o palabra en RAH) tiene más probabilidad de ocurrencia en un tiempo determinado. Dado un autómata probabilístico, cada estado puede representar o emitir un evento. De esta forma se puede tener un modelo con las probabilidades de que evento puede continuar de los eventos disponibles. Esta estructura es conocida por el nombre de *gramática* y en la Figura 2.5 se muestra la idea general de la misma. La gramática es un conjunto de reglas que limitan el número de combinaciones de eventos permitidos a fin de mejorar la tasa de reconocimiento y eliminar ambigüedades. De esta manera, se sitúa el estudio por encima de las características acústicas.

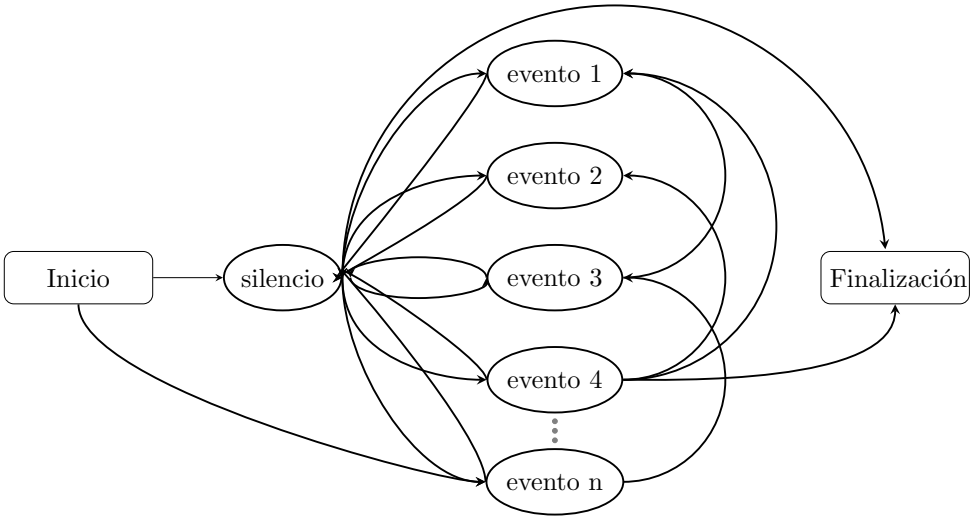

Figura 2.5. Modelo de lenguaje

Capítulo 3

Corpus de emociones

A pesar de haber sido estudiado desde los años 50, la investigación de las señales emocionales en el discurso está ganando cada vez mayor atención. Esto se debe principalmente al auge de interfaces hombre-máquina, que ven las aplicaciones de reconocimiento automático y simulación de la expresión emocional dentro de su alcance.

En este capítulo se describe la planificación y realización de dos bases de habla emocional empleadas para los experimentos con el reconocedor. Se utilizó una base de datos en alemán que fue desarrollada por un grupo de investigadores con actores alemanes. La misma fue elegida por estar conformada por múltiples hablantes, lo que permite evaluar la eficiencia del reconocedor de forma independiente del hablante. Esta es una característica muy importante a la hora de validar los métodos usados para desarrollar el reconocedor. Como se detalló en la Sección 1.2, en estudios previos se han obtenido muy buenos resultados pero con la utilización de bases de datos conformadas por uno o dos hablantes, lo que los hace inaceptables desde un punto de vista práctico. Debido a la falta de una base de datos de habla emocional en español de Argentina, se elaboró un pequeño corpus, a partir de la extracción de audios de películas argentinas.

En la primera parte del Capítulo se argumenta por qué el material fue grabado con voces actuadas en lugar de voces en situaciones de la "vida real". La segunda sección describe la base de datos en alemán a partir del documento [34] redactado por los creadores de la base de datos. Se explica el objetivo de las emociones que fueron elegidas, la elección de los materiales de texto, seguido del registro, la evaluación y el etiquetado de los datos. Por último se explica detalladamente los pasos que se siguieron en el desarrollo de la base de datos emocional en español.

3.1. ¿Emociones actuadas o reales?

Existen muchos argumentos en contra de la expresión emocional actuada. Como se señala en [35], las emociones auténticas muy rara vez aparecen en el mundo real. Además, existen señales físicas que no pueden ser imitadas conscientemente. Sin embargo, la más clara expresión emocional no sólo es poco frecuente en situaciones de la vida cotidiana, sino también el registro de las personas que experimentan emociones auténticas. Por esto, es casi imposible la generación de base de datos de habla espontánea si las emociones son objeto de exploración.

Para encontrar un punto de inflexión entre estas dos realidades, se pueden tomar medidas que permitan que la emoción no sea del todo actuada ni del todo natural. Una medida, consiste en aflorar la emoción a partir de recuerdos de la propia vida del actor. Este método es conocido por *método Stanislavski* y fue aplicado en la creación de la base de datos en alemán.

3.2. Base de datos de habla emocional en alemán

Esta base de datos de habla emocional en alemán ha sido desarrollada por el Instituto de Ciencias de la Comunicación de la Technical University de Berlin [34]. Esta base de datos ha sido utilizada para numerosos estudios (entre los cuales se pueden citar [36], [37] y [38]) y está disponible gratuitamente en Internet[1]. A continuación se describen los detalles de esta base de datos siguiendo la presentación [34].

3.2.1. Generalidades

Esta base de datos incluye seis emociones además del tipo neutral. Las emociones son: alegría, ansiedad, susto, cansancio, enojo y tristeza. Las grabaciones fueron realizadas por diez actores, cinco masculinos y cinco femeninos, que simularon un corpus de frases y párrafos expresando las distintas emociones. Para cada tipo de emoción se dispone de grabaciones de 1 a 7 segundos de duración. El total de señales de voz asciende a 535, con la distribución que se ve en la Tabla 3.1.

Para la realización del corpus, se fijaron los siguientes objetivos:

- Un número razonable de oradores deben generar todas las emociones para ofrecer independencia del hablante.

[1] La información es accesible desde http://pascal.kgw.tu-berlin.de/emodb/.

Tabla 3.1. Distribución del corpus de emociones en alemán.

Emoción	Cantidad
Alegría	71
Ansiedad	69
Susto	46
Cansancio	62
Enojo	81
Tristeza	127
Neutral	79

- Todos los oradores deben pronunciar el mismo contenido verbal con el fin de permitir la comparación entre las emociones y los oradores.
- Las grabaciones de los audios deben ser de alta calidad, minimizando el ruido de fondo.
- Utilizar una cámara anecoica con el fin de anular los efectos de eco y reverberación del sonido, y un laringógrafo para obtener las características de la señal de habla en el momento en que el hablante la expresa.

Para cumplir con los objetivos se incorporan frases cortas donde puede expresarse cada emoción. Las frases utilizadas por los actores se presentan en el Apéndice A.

3.2.2. La elección de los actores

Los actores profesionales tienden a expresar las emociones de manera exagerada. Esto se le presentaba como una dificultad al momento de la elección de las personas que expresarían las emociones. La solución consistió en buscar los oradores a través de un aviso en el diario. Al mismo respondieron 40 personas, a las que se les pidió que expresen las emociones en una oficina que contaba con un micrófono. Luego, tres oyentes expertos seleccionaron 10 personas, 5 femeninos y 5 masculinos a partir de la naturalidad de las expresiones de las emociones. De todos los seleccionados, excepcionalmente uno había realizado estudios de actuación.

3.2.3. Grabación de los audios

Para lograr una alta calidad de audio de la grabación se utilizó una cámara anecoica en la Universidad Técnica de Berlín, Departamento Técnico

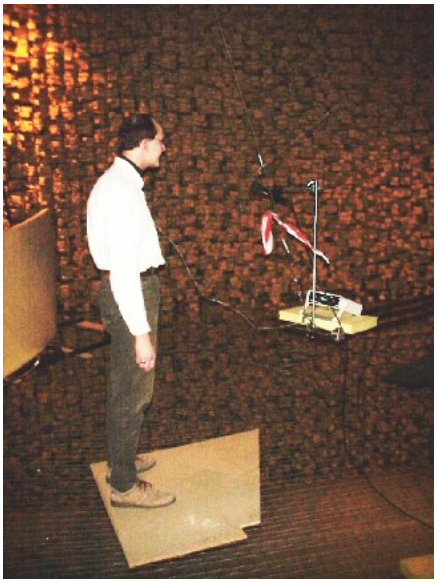

Figura 3.1. Un actor durante la grabación en la cámara anecoica.

de Acústica usando un micrófono Sennheiser MKH 40 P 48 y una Tascam DA-P1 grabadora DAT portátil. Además se utilizó un laringógrafo portátil. Las señales fueron grabadas a una frecuencia de muestreo de 48kHz y luego remuestradas a 16kHz.

Los actores estaban de pie delante del micrófono para que pudieran utilizar el lenguaje corporal y sólo les entorpecía el cable del laringógrafo y la necesidad de posicionarse frente al micrófono a una distancia de unos 30 cm. (véase la Figura 3.1 extraída de [34]). La grabación de cada participante se llevó a cabo en una única sesión que duró aproximadamente dos horas. Al elegir la emoción se caracterizaba la misma a través de relatos y con tiempo razonable para situarse en esa emoción. A su vez se les pidió a los actores que recuerden una situación real de su pasado cuando habían sentido esa emoción. Los actores podían decir las frases en la frecuencia que querían por lo que en algunos casos se registró mas de una versión.

3.2.4. Test perceptual

Para garantizar la calidad y naturalidad emocional se llevó a cabo un test de percepción en el que participaron 20 personas. Se les permitió escuchar cada grabación sólo una vez y luego decidir en qué estado emocional

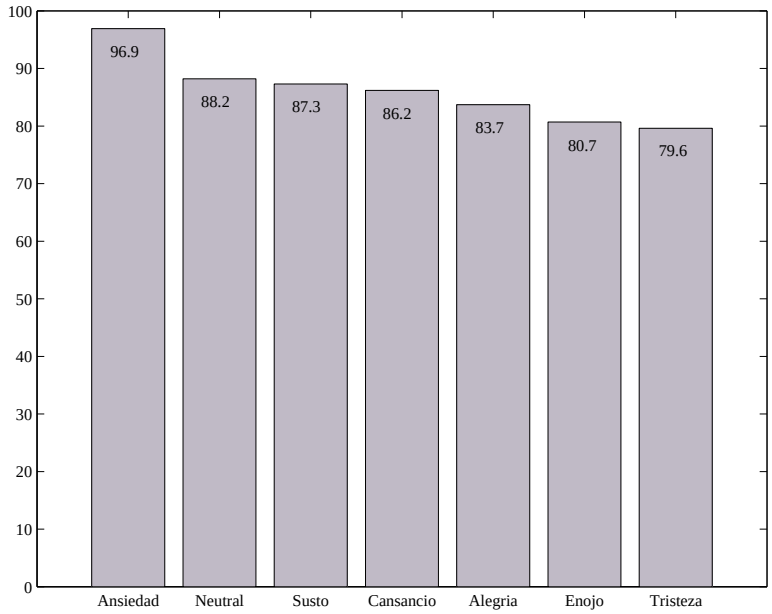

Figura 3.2. Resultados del test de percepción con la base de datos en alemán.

se había dicho la elocución. Posteriormente se seleccionaron los audios que obtuvieron una tasa de reconocimiento mayor al 80 % y una naturalidad mayor al 60 %. En la Figura 3.2 se muestra la tasa de reconocimiento para las siete emociones involucradas.

3.3. Base de datos de habla emocional en español

3.3.1. Generalidades

La motivación principal que llevó a la producción de un corpus de habla expresiva en español de Argentina fue la falta de disponibilidad de un recurso de este tipo, que permitiese evaluar el desempeño del reconocedor en nuestra lengua materna. El número de bases de datos existentes en el ámbito nacional e internacional es reducido y con características que no responden a esta necesidad.

El primer paso en la producción del corpus de emociones consistió en planificar las tareas y determinar los elementos necesarios para obtener los audios que integran la base de datos en español. Las características que se

Tabla 3.2. Distribución de los audios extraídos de películas en español.

Emoción	Cantidad
Alegría	40
Enojo	74
Neutral	62

pudieron lograr en el corpus respetaron una serie de objetivos, a pesar de las limitaciones propias de las películas que se utilizaron para su creación. Las realizaciones de este tipo de corpus, puede necesitar varios años de trabajo para poder obtener una base de datos de audios con las características necesarias para validar sistemas de reconocimiento de emociones. Aquí, se marcó un objetivo menos ambicioso y coherente con la extensión del proyecto final de carrera, pero que permitiera ser la base para experimentar con el reconocedor en español, con un costo económico y un tiempo de desarrollo que fuera posible asumir. Por estas razones se fijan los siguientes objetivos:

- Se consideran 3 tipos de emociones: alegría, enojo y neutral.
- Los fragmentos de audio se extraen de películas argentinas.
- Los actores considerados serán nativos de Argentina y deben expresar todas las emociones.
- Por cada actor se extraen diez audios en los que exprese cada tipo de emoción, teniendo como resultado final 30 audios del mismo orador.

Por cada actor se extrajeron alrededor de 15 fragmentos de audio para cada tipo de emoción, luego se descartaron aquellos que presentaron una baja tasa de reconocimiento en el test de percepción (ver Sección 3.3.4). En la Tabla 3.2 se observa la distribución de los audios que se extrajeron en función del tipo de emoción. A diferencia de la base de datos en alemán, las frases expresadas por los actores no son iguales debido a que los fragmentos de audios se extrajeron de películas. Las transcripciones han sido ignoradas por ser irrelevantes para los métodos de clasificación propuestos.

3.3.2. La elección de los actores

La elección de los actores se basó en los objetivos previamente expuestos. Se seleccionaron dos actores y dos actrices, que participaron en varias

Tabla 3.3. Distribución del corpus de emociones en español utilizado en el reconocedor de emociones.

Emoción	Cantidad
Alegría	36
Enojo	40
Neutral	40

películas argentinas, las cuales abarcan diversas temáticas y tienen situaciones que permiten encontrar diálogos en los que se expresan las emociones buscadas. Hoy en día, muchas producciones de películas son españolas-argentinas por lo que al momento de la elección de los actores se eligió aquellos que sean hablantes nativos de Argentina y que no tengan un acento regional muy marcado. Los actores que se tomaron para la extracción de los audios fueron: Jorge Marrale, Diego Peretti, Carolina Peleritti y Cecilia Roth.

3.3.3. Extracción de los segmentos de audio

Se seleccionaron 4 actores y se extrajeron expresiones emocionales para cada actor, con una duración de entre 3 y 5 segundos. Cada audio fue normalizado a una amplitud máxima de -3 dB, remuestreados a una frecuencia de muestreo de 48 kHz, en un solo canal. Las películas de las cuales se extrajeron los audios se listan en el apéndice B.

3.3.4. Test perceptual

Al igual que para la base de datos en alemán, se realizó un test perceptual en el que participaron 20 personas. Para esto, se le proporcionó a cada persona todos los audios dispuestos de forma aleatoria. Se confeccionó una planilla para los evaluadores donde cada audio se identificaba con un número y la secuencia de audios estaba desordenada respecto de las emociones y de los actores. A partir de las planillas, se procedió a obtener el porcentaje de reconocimiento de cada tipo de emoción para cada uno de los audios. En la Figura 3.3 se pueden observar el porcentaje de reconocimiento en función del tipo de emoción.

Se puede observar que las emociones son claramente identificadas por los participantes del test, siendo la neutral el tipo de emoción reconocido con más exactitud. El promedio de aciertos considerando la totalidad de las personas encuestadas es de 90.40 %. En la Tabla 3.3 se observa la distribución definitiva de los audios en función del tipo de emoción.

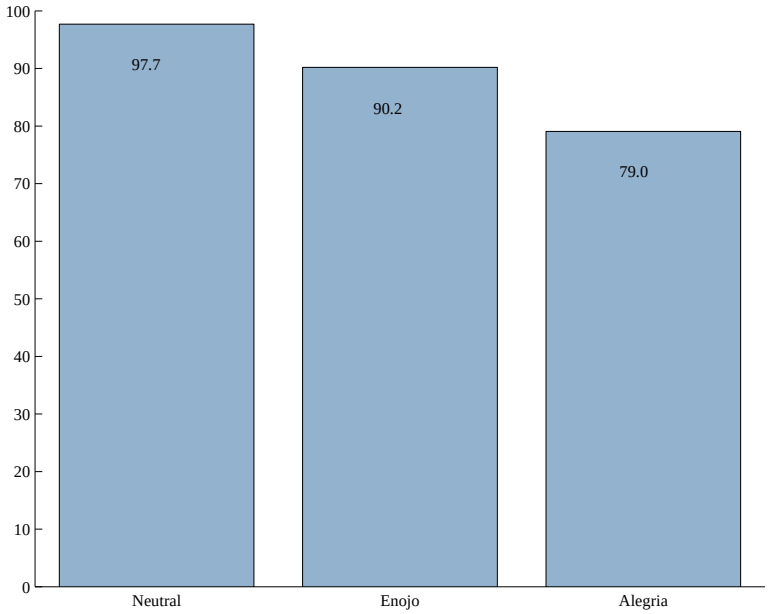

Figura 3.3. Resultados del test de percepción con la base de datos en español.

Debido a que los audios se extrajeron de películas, se presentaron diferentes dificultades como la falta de control sobre el contenido y habitualmente, una calidad de audio insuficiente debido a las condiciones de grabación. En muchas ocasiones los audios poseen cierto contenido emocional mezclado con una banda sonora para enfatizar la situación, lo que convierte en inutilizable al audio extraído. Otro problema frecuente fue la interposición de voces.

Capítulo 4

Diseño e implementación del sistema

Una vez desarrollada la base de datos en español y obtenida correctamente la base de datos en alemán, se procedió al diseño e implementación de un reconocedor automático de emociones. En este capítulo se detallan los pasos que siguen para el desarrollo del sistema de reconocimiento de emociones. A partir de los conceptos explicados en el Capítulo 2, se describen el método empleado para la extracción de características de los audios y la definición de los modelos. Para esta última etapa, se definen prototipos de dos modelos: mezcla de gaussianas y modelos ocultos de Markov. Después de definir los prototipos, se sigue con las fases de entrenamiento y prueba del reconocedor. Por último se detallan las modificaciones que se realizaron en los modelos y las evaluaciones de sus resultados, con el fin de obtener el reconocedor más eficiente tanto para el caso del modelo basado en HMM como para el modelo basado en GMM.

4.1. Extracción de características

Para extraer las características de la señal de voz se utilizan herramientas del procesamiento digital de señales. Éste no es un paso menor en el desarrollo del reconocedor, ya que al analizar en detalle la señal del habla se obtienen las características de mayor relevancia de la voz, dando lugar a mejorar el rendimiento de un sistema de reconocimiento automático de emociones. En esta etapa, se realiza el análisis por tramos y la parametrización de la señal de voz.

Para el análisis por tramo se utilizó la ventana de Hamming ya que

resulta ser de las más indicadas para el análisis de señales de voz [26]. La señal se segmentó en tramos sucesivos, cuya longitud es de 25 ms solapados en 10 ms.

Se obtuvo un conjunto de vectores con las propiedades espectrales de los audios, siguiendo la siguiente metodología: se parametrizaron los audios a través de MFCC (Coeficientes Ceptrales en la escala de Mel). Se obtuvieron los primeros 12 coeficientes MFCC, 12 "coeficientes delta" y 12 "coeficientes de aceleración". Esto da un vector de 36 coeficientes en cada ventana de tiempo de análisis. Para los audios de la base de datos en español se adicionó el coeficiente de la energía con sus coeficientes delta y aceleración, ascendiendo a 39 el largo del vector en cada ventana temporal de análisis. El filtro de pre-énfasis que se aplicó a cada una de las señales, es de tipo pasa alto con una estructura $1 - a_1 z^{-1}$ donde el coeficiente es $a_1 = 0{,}97$ [24].

El resultado en esta etapa es un archivo de coeficientes MFCC, para cada ventana, por cada audio de la base de datos.

4.2. Definición del modelo de referencia

Para el desarrollo e implementación de los modelos de GMM y HMM se utilizó un conjunto de herramientas denominado Hidden Markov Toolkit [39]. El mismo permite variar fácilmente la definición de los modelos y los parámetros de entrenamiento, por lo que se propuso utilizarlo en el proceso de construcción del reconocedor automático de emociones a través de la voz. En esta etapa se define, por un lado, el reconocedor de emociones basado en GMM y por el otro el reconocedor de emociones basado en HMM.

Para el diseño de GMM se definió un modelo compuesto por una cantidad n de gaussianas en la mezcla. Para estimar el número óptimo de gaussianas para el modelo se han desarrollado GMMs de diferente orden. El proceso de incremento de la mezcla se realiza adicionando una pequeña cantidad de gaussianas por vez. De esta forma se puede determinar cuando un modelo compuesto por una mezcla de gaussianas es óptimo.

Para diseñar un modelo oculto de Markov se deben tener en cuenta los siguientes puntos:

- La cantidad de estados del modelo.
- Las distribuciones de probabilidad de observación.
- Las transiciones entre estados.

La forma de transición entre estados es la denominada de *izquierda a derecha* y es la que se muestra en la Figura 2.2. En cada estado del modelo se puede observar con una mezcla de gaussianas, por lo tanto, las *medias* y las *varianzas* son los parámetros estadísticos seleccionados para modelar la distribución de probabilidades de las observaciones de los estados. A partir de estas definiciones generales, la topología del HMM se ha alterado a lo largo de los experimentos. Por ejemplo: se incrementó la cantidad de estados del modelo; se ajustó la matriz de probabilidades de transición a_{ij}; se incrementó la cantidad de gaussianas en las mezclas.

La gramática que se definió para el reconocedor es simple: un silencio inicial, seguida de una emoción y luego un silencio final. La red gramatical utilizada en el reconocedor de emociones de este proyecto se muestra en la Figura 4.1.

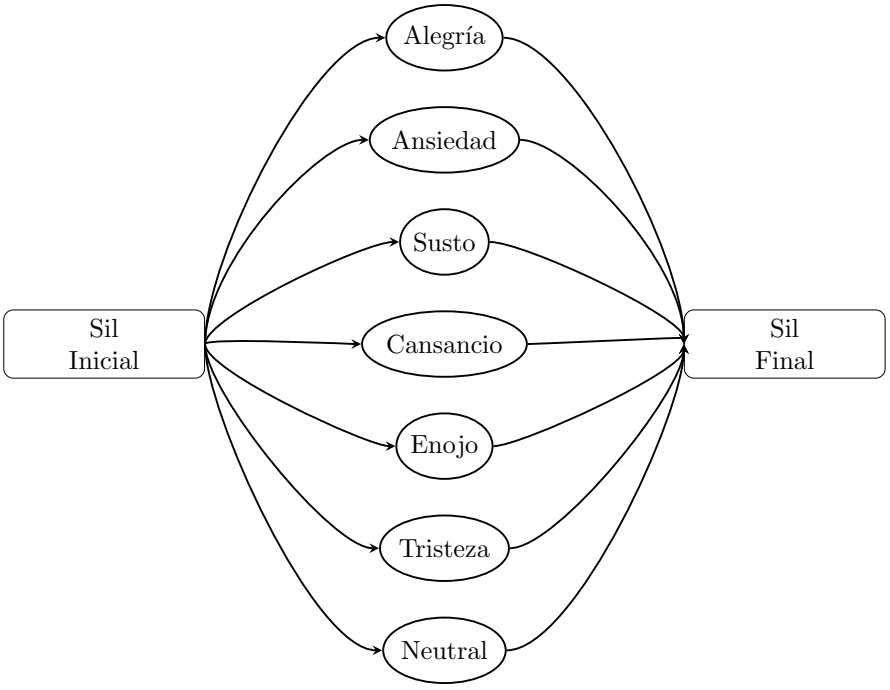

Figura 4.1. Red gramatical

En un archivo de texto, se define una tabla de todas las emociones que pueden ser reconocidas y la estructura del modelo HMM asociado a la

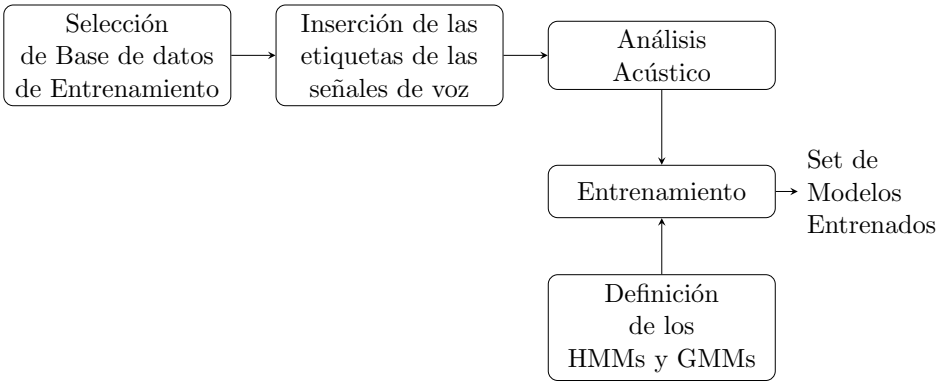

Figura 4.2. Diagrama de bloques: fase de entrenamiento

emoción. Además se define el silencio inicial y final los cuales hacen referencia al mismo modelo HMM. Este archivo se lo conoce como el diccionario del reconocedor.

4.3. Entrenamiento del modelo

En esta etapa, el modelo recibe como entrada los vectores de características previamente obtenidos y se ajustan todos lo parámetros de cada modelo HMM para que represente globalmente a todas las secuencias observadas para esta clase. En el diagrama de la Figura 4.2, se pueden observar las etapas involucradas en la fase de entrenamiento del modelo.

El procedimiento de entrenamiento está descrito en la Figura 4.3, tanto para el modelo basado en GMM como para el modelo basado en HMM.

Al entrenar los modelos, se realiza una reestimación de sus parámetros, los cuales se van guardando en un directorio. A partir del modelo inicial, se entrena hasta llegar a la convergencia. En el transcurso del entrenamiento del modelo se ejecuta una re-estimación de parámetros por Baum-Welch [24] y se itera hasta obtener los parámetros óptimos del mismo (probabilidades de transición; y medias y varianzas para cada distribución de observación).

En el caso de entrenamiento de GMM, se comenzó con una cantidad de gaussianas en la mezcla y se fué incrementando en cada iteración. Cada vez que se añaden gaussianas a la mezcla se obtiene un modelo que luego se ajusta realizando una cantidad de reestimaciones de Baum-Welch. El

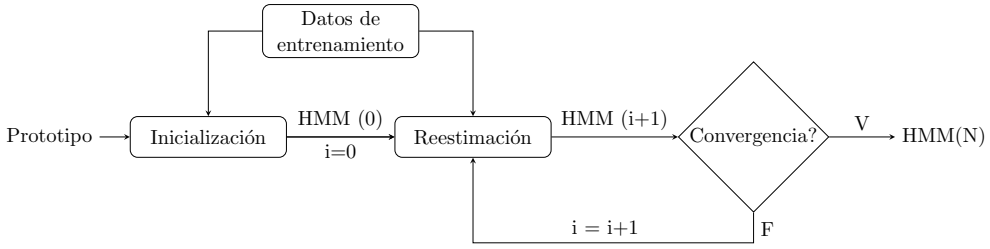

Figura 4.3. Entrenamiento de un GMM/HMM

procedimiento de entrenamiento se realiza hasta alcanzar el número deseado de gaussianas en la mezcla.

En el caso del entrenamiento de los HMM, se procedió del mismo modo, con la distinción de que el incremento de gaussianas en la mezcla se produce en cada estado del modelo.

4.4. Prueba del modelo

El rendimiento del sistema puede ser mal estimado si se utiliza una única partición de entrenamiento y prueba. Se puede producir una tendencia en el error de reconocimiento, a favor o en contra, ocasionada por la selección peculiar de las frases en cada uno de los conjuntos. Para evitar esto, las pruebas se realizaron por validación cruzada [40] según el método de *averaged leave-k-out*. Se generaron 10 particiones de datos, para cada una se tomó de forma aleatoria el 80 % de los datos para el entrenamiento y el 20 % restante para la etapa de prueba.

En la fase de reconocimiento o prueba, en la cual se utilizan los audios de la base de datos que no han sido utilizados en la fase de entrenamiento se evalúa el rendimiento del reconocedor. Como datos de entrada, se tiene: el diccionario, la red gramatical, los modelos entrenados y los audios parametrizados. La salida de esta fase son las etiquetas de las emociones reconocidas por el sistema, las cuales se comparan con las etiquetas originales y se obtienen las estadísticas del reconocedor. En la Figura 4.4 se puede observar el proceso descrito.

Un sistema de reconocimiento se evalúa comparando las etiquetas correctas y las reconocidas. Para analizar la eficiencia del reconocedor se calcula el porcentaje de aciertos, como

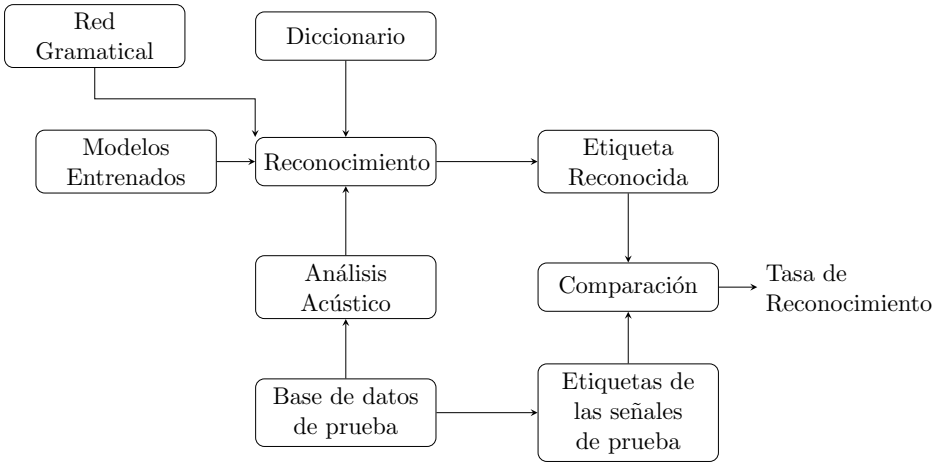

Figura 4.4. Diagrama de bloques: fase de reconocimiento

$$\%Corr = \frac{H}{N} \times 100\,\% \tag{4.1}$$

donde H es la cantidad de aciertos del reconocedor y N es la cantidad total de audios evaluados. El valor de H se determina a partir de los errores de sustitución de una emoción por otra.

4.5. Alternativas de diseño evaluadas

Con el objetivo de lograr sistemas más eficientes dada una cantidad de emociones, se procedió a ajustar diferentes variables como la parametrización, el número de gaussianas en el modelo GMM, la cantidad de estados en los modelos HMM, la cantidad de reestimaciones, la cantidad de gaussianas en las mezclas de los estados del HMM.

Con la base de datos en español se hicieron experimentos para evaluar la importancia del coeficiente de energía, comparando los resultados de considerarlo o no en la parametrización.

Se realizaron las pruebas con GMM variando el número de componentes gaussianas en la mezcla, incrementando de a 2 cada vez hasta llegar a un modelo con 512 componentes gaussianas. Como la cantidad de gaussianas en la mezcla no presentó una variación significativa a partir de las 32 gaussianas, se definió esta cantidad como máxima.

Para la evaluación de los modelos ocultos de Markov se definió un modelo de 2 estados incrementando el número de gaussianas del mismo modo que las pruebas del modelo GMM pero con la distinción de que las gaussianas se adicionan a la mezcla de cada estado del modelo. Luego, se realizó otro experimento agregando un estado más al modelo, así sucesivamente hasta realizar un experimento con 7 estados. En cada caso se debió reestimar los parámetros del modelo.

Con cada base de audios se evaluó tanto el modelo de GMM como los HMM de 2, 3, 4, 5, 6 y 7 estados. Con la base de audios en alemán, se evaluaron distintas cantidades de emociones, comenzando por tres (enojo, alegría y neutral) e incrementando en una emoción hasta incluir las siete emociones.

También se evaluó la cantidad de reestimaciones necesarias del algoritmo de Baum-Welch luego del incremento de gaussianas al modelo. Las reestimaciones se realizaron mientras hubo un cambio significativo del modelo entre una iteración y otra, de lo contrario se finalizó el entrenamiento. Se encontró que no es necesario hacer más de seis reestimaciones pues de los experimentos resulta que esta cantidad es adecuada.

Capítulo 5

Evaluación de resultados

En este capítulo se presentan las pruebas más importantes realizadas en este trabajo. Se evaluan y comparan las distintas técnicas de reconocimiento de emociones descritas en los capítulos anteriores.

El presente capítulo se divide en dos subsecciones. En la primera se presentan y analizan los resultados obtenidos en los experimentos realizados con la base de datos en alemán, mientras que en la segunda parte se presentan los obtenidos con la base de datos en español. Como se ha mencionado, la implementación del sistema de reconocimiento de emociones se realizó mediante las técnicas de clasificación GMM y HMM.

Para los experimentos, se ajustaron las variables según la cantidad de emociones y el método a utilizar, para luego proceder al entrenamiento del clasificador. Como se dijo anteriormente, para medir la exactitud del reconocedor se utilizó el método validación cruzada [11], empleando 10 particiones. Los audios de cada partición se eligieron al azar. Cada partición se dividió en dos, un 80 % de los datos para entrenamiento y un 20 % para prueba. Los resultados que se presentan corresponden al promedio de las 10 particiones.

Los errores en estos experimentos se deben a la sustitución de una emoción por otra. La matriz de confusión es una buena forma de analizar qué modelos tuvieron errores de reconocimiento y cómo se presentaron. En esta matriz las entradas al reconocedor se disponen en las filas y en las columnas se ubican las salidas obtenidas por el reconocedor. La diagonal principal de la matriz presenta el número de emociones que fueron reconocidas correctamente, es decir, los aciertos del reconocedor. Fuera de la diagonal principal se observan los errores, lo que permite apreciar con mayor claridad cuales son las clases mas difíciles de clasificar.

5.1. Resultados con el corpus de emociones en alemán

5.1.1. Reconocimiento de emociones con GMM

Se comenzó el análisis utilizando 3 emociones: alegría, enojo y neutral. La Tabla 5.1 muestra la matriz de confusión para el reconocimiento de éstas emociones con el método GMM. Con 22 componentes gaussianas se ha logrado un reconocimiento del 79 %, con las confusiones más importantes entre las emociones de alegría y enojo.

Se realizó otro experimento donde se incorpora una emoción más: la *tristeza*. En la Tabla 5.2 se pueden apreciar las confusiones que se producen entre las emociones. La tasa de aciertos en este caso fue 68.43 % con un modelo de 26 gaussianas. La incorporación de la tristeza, produjo mucha confusión entre ésta y el estado neutral.

La incorporación de una emoción más (el susto), hizo necesaria una mayor cantidad de gaussianas para poder obtener mejores modelos para la clasificación. En la Tabla 5.3 se aprecia la importante confusión entre: alegría y enojo, tristeza y neutral. La incorporación del *susto* en el experimento, produjo que el reconocedor la confunda en algunos casos con la alegría y en menos ocasiones con las restantes emociones. La tasa de acierto en este experimento fue del 71.39 % con un modelo de 32 gaussianas.

En un nuevo experimento, para 6 emociones, se adicionó la emoción *ansiedad*. En este caso, las pruebas arrojaron un porcentaje del acierto del 68.80 % con un modelo de 16 gaussianas. La nueva emoción, se confunde principalmente con la alegría y neutral. En la Tabla 5.4 puede observarse la matriz de confusión del experimento para 6 emociones con el modelo GMM.

En la Tabla 5.5 se muestra una matriz de confusión para el reconocimiento de emociones con 32 componentes gaussianas y siete emociones. Allí, el porcentaje de las emociones correctamente identificadas fue del 67.40 %. La última emoción en incorporarse en las pruebas fue *cansancio* y se puede apreciar que el reconocedor sólo tuvo, para ésta, un alto porcentaje de confusión con la tristeza.

5.1.2. Reconocimiento de emociones con HMM

En la Figura 5.1 se muestran las tasas de reconocimiento obtenidas por el sistema de reconocimiento de emociones con el corpus de emociones en alemán sobre las diez particiones. Los porcentajes obtenidos son un promedio de los resultados de los modelos para todas las cantidades de componentes gaussianas (de 1 a 32), manteniendo fijo la cantidad de estados y la cantidad

Tabla 5.1. Matriz de confusión para 3 emociones con GMM (22 gaussianas)[BD en alemán]

Emoción	Alegría	Enojo	Neutral
Alegría	**99**	38	3
Enojo	50	**192**	8
Neutral	11	4	**135**

Tabla 5.2. Matriz de confusión para 4 emociones con GMM (26 gaussianas)[BD en alemán]

Emoción	Alegría	Enojo	Tristeza	Neutral
Alegría	**83**	41	1	15
Enojo	36	**203**	2	9
Tristeza	3	3	**108**	46
Neutral	10	4	51	**85**

Tabla 5.3. Matriz de confusión para 5 emociones con GMM (32 gaussianas)[BD en alemán]

Emoción	Alegría	Susto	Enojo	Tristeza	Neutral
Alegría	**91**	13	31	3	2
Susto	11	**67**	5	3	4
Enojo	32	0	**213**	0	5
Tristeza	4	18	2	**90**	46
Neutral	5	9	9	24	**103**

Tabla 5.4. Matriz de confusión para 6 emociones con GMM (16 gaussianas)[BD en alemán]

Emoción	Alegría	Ansiedad	Susto	Enojo	Tristeza	Neutral
Alegría	**75**	20	5	40	0	0
Ansiedad	13	**85**	4	9	6	13
Susto	3	8	**65**	4	7	3
Enojo	35	3	3	**209**	0	0
Tristeza	1	4	8	0	**108**	39
Neutral	4	13	7	7	28	**91**

Tabla 5.5. Matriz de confusión para 7 emociones con GMM (32 gaussianas)[BD en alemán]

Emoción	Alegría	Ansiedad	Susto	Cansancio	Enojo	Tristeza	Neutral
Alegría	**101**	9	2	0	25	0	0
Ansiedad	21	**62**	2	16	17	4	4
Susto	4	12	**67**	8	3	3	0
Cansancio	0	0	1	**100**	0	14	5
Enojo	23	6	1	0	**220**	0	0
Tristeza	0	6	15	26	0	**63**	50
Neutral	2	4	6	21	3	30	**84**

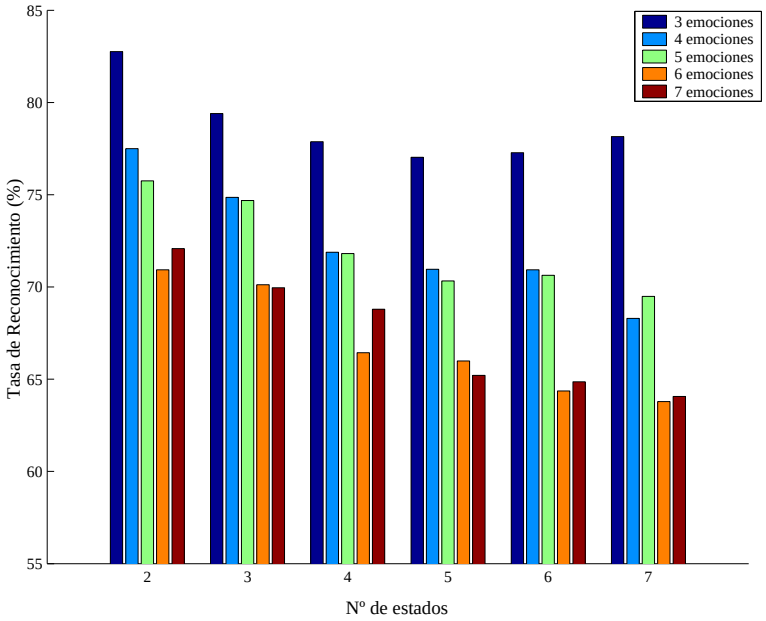

Figura 5.1. Reconocimiento de emociones en función de la cantidad de estados utilizando HMM [BD en alemán]

de emociones. Pueden observarse los rendimientos del reconocedor según se varía el número de estados del modelo HMM y el número de emociones diferentes que se reconocen. Es posible observar que no hay necesidad de aumentar el número de estados más allá de dos. Por lo tanto, los resultados que se muestran a continuación corresponden a un reconocedor de emociones aplicando la técnica de modelos ocultos de Markov con 2 estados.

El desempeño del reconocimiento de HMM también fue evaluado en relación al número de componentes gaussianas de sus estados. En la gráfica de la Figura 5.2 se compara la tasa de reconocimiento en función de la cantidad de gaussianas. Se aprecia la evolución del reconocimiento al incrementar el número de gaussianas en las mezclas, en experimentos con modelos de dos estados y variando la cantidad de emociones. En el rango de 14 a 22 componentes gaussianas es donde el rendimiento deja de incrementarse significativamente.

El reconocedor de emociones basado en un clasificador de modelos ocultos de Markov de 2 estados, fue evaluado del mismo modo que el clasificador GMM. La Tabla 5.6 muestra la matriz de confusión obtenida al reconocer 3 emociones. Para el modelo de 2 estados y 14 componentes gaussianas, se obtuvo una tasa de acierto del 86 %. Se puede ver una importante confusión

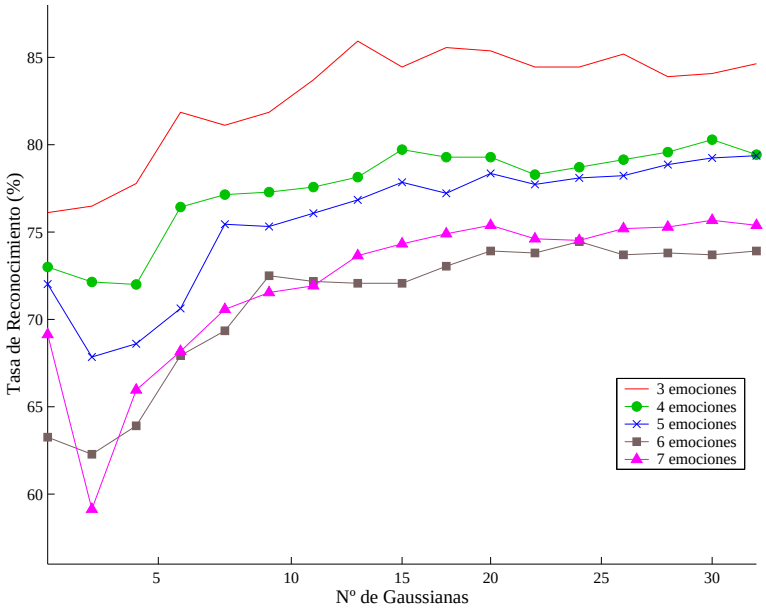

Figura 5.2. Reconocimiento de emociones en función de la cantidad de gaussianas, con modelos HMM de 2 estados [BD en alemán]

entre alegría y enojo.

En el experimento con 4 emociones, las emociones involucradas fueron: alegría, enojo, tristeza y neutral. El resultado obtenido en dicha prueba muestra un 80.29 % de emociones reconocidas correctamente con 30 gaussianas en las mezclas de los estados. En la Tabla 5.7 se muestra la matriz de confusión obtenida del proceso de reconocimiento. Este reconocedor produjo confusiones similares a las del experimento visualizado en la Tabla 5.2, pero con un porcentaje menor de error.

La siguiente emoción incorporada fue el *susto*. Al evaluar el desempeño del reconocedor, se obtuvo un porcentaje de aciertos del 79.37 % con 32 componentes gaussianas. La matriz de confusión de esta prueba se aprecia en la Tabla 5.8. Se mantienen sustituciones entre emociones similares a los experimentos anteriores, y la nueva emoción tiene mayor confusión con la emoción de tristeza.

El reconocedor para 6 emociones obtuvo una tasa de reconocimiento de 74.46 % en un modelo con 24 componentes gaussianas por estado. En esta instancia se añadió la emoción de *ansiedad*, y en la Tabla 5.9 se observa la matriz de confusión del experimento.

En la Tabla 5.10 se puede observar el resultado obtenido al utilizar

Tabla 5.6. Matriz de confusión para 3 emociones con HMM (2 estados, 14 gaussianas)[BD en alemán]

Emoción	Alegría	Enojo	Neutral
Alegría	**101**	39	0
Enojo	32	**216**	2
Neutral	2	1	**147**

Tabla 5.7. Matriz de confusión para 4 emociones con HMM (2 estados, 30 gaussianas)[BD en alemán]

Emoción	Alegría	Enojo	Tristeza	Neutral
Alegría	**104**	35	0	1
Enojo	24	**224**	0	2
Tristeza	4	0	**113**	43
Neutral	0	0	29	**121**

Tabla 5.8. Matriz de confusión para 5 emociones con HMM (2 estados, 32 gaussianas)[BD en alemán]

Emoción	Alegría	Susto	Enojo	Tristeza	Neutral
Alegría	**98**	6	35	0	1
Susto	7	**65**	2	14	2
Enojo	18	0	**230**	0	2
Tristeza	0	6	14	**110**	30
Neutral	2	5	2	27	**114**

Tabla 5.9. Matriz de confusión para 6 emociones con HMM (2 estados, 24 gaussianas)[BD en alemán]

Emoción	Alegría	Ansiedad	Susto	Enojo	Tristeza	Neutral
Alegría	**83**	16	1	40	0	0
Ansiedad	9	**90**	4	9	11	7
Susto	6	10	**65**	1	6	2
Enojo	24	11	1	**213**	0	1
Tristeza	0	7	7	0	**122**	24
Neutral	0	6	5	0	27	**112**

Tabla 5.10. Matriz de confusión para 7 emociones con HMM (2 estados, 30 gaussianas)[BD en alemán]

Emoción	Alegría	Ansiedad	Susto	Cansancio	Enojo	Tristeza	Neutral
Alegría	**93**	13	0	0	34	0	0
Ansiedad	13	**93**	5	6	7	3	3
Susto	4	7	**70**	0	4	3	2
Cansancio	0	0	3	**93**	0	23	1
Enojo	12	6	1	0	**231**	0	0
Tristeza	0	3	7	14	0	**94**	42
Neutral	0	5	5	0	0	27	**113**

las siete emociones que componen la base de datos. En este caso la tasa de acierto alcanzó el 76 % en un modelo con dos estados HMM y 30 componentes gaussianas. Se puede observar que la tasa de reconocimiento es mayor a la obtenida al utilizar el método GMM, lo que da cuenta de la utilidad de los HMM. La emoción que mejor se clasifica es el enojo (92.4 %), seguida del susto (77.7 %) y el cansancio (77.5 %). Sin embargo, el reconocimiento de la tristeza es baja (54.75 %). Si analizamos en detalle la matriz de confusión, podemos dividir las emociones en dos grandes grupos, atendiendo a la similitud que existe entre ellas según su grado de confusión. Así, la alegría, el enojo y el susto, emociones con un alto nivel de activación, se confunden principalmente entre ellas. Por el contrario, emociones negativas o pasivas como la neutra, la tristeza o el cansancio se confunden entre sí; encontrándose la ansiedad en medio de estas dos agrupaciones.

La primera observación acerca de los resultados es el hecho de que en todos los casos el corpus utilizado esta conformado por las emociones de 10 personas, 5 mujeres y 5 hombres. La independencia de locutor es un aspecto determinante para el correcto funcionamiento de un reconocedor de emociones que esté destinado a distintos ambientes y múltiples hablantes. La implementación de reconocedores dependientes de locutor tiene sentido cuando se está hablando de aplicaciones que serían usadas por un único locutor y se requiera un grado de reconocimiento elevado. Sin embargo, no es adecuado para la mayoría de las aplicaciones prácticas. Los resultados obtenidos en el presente proyecto no son dependientes del locutor ya que el corpus empleado está constituido por diferentes hablantes.

En principio las emociones: enojo, alegría y neutral, se podrían considerar emociones no relacionadas e instintivamente proponerlas como extremas para la clasificación. Sin embargo, las pruebas con otro grupo de emociones (enojo, ansiedad y neutral) han obtenido mejores resultados, un 97 % de acierto. Este resultado sugiere que estas últimas son emociones más diferenciables cuando son expresadas en la voz, lo que podría dar lugar a análisis acústicos más profundos y, en base a éstos, a una posible redefinición de lo que comúnmente se consideran emociones primarias y secundarias.

5.2. Resultados con el corpus de emociones en español

Con el fin de comprobar la eficacia del reconocedor en nuestro idioma, se procedió a realizar pruebas similares a las que se realizaron con la base de datos en alemán. Cabe destacar que el objetivo principal es estimar la

Tabla 5.11. Matriz de confusión para 3 emociones con GMM (18 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**44**	19	7
Enojo	10	**65**	5
Neutral	23	22	**35**

adaptabilidad y la eficiencia del reconocedor, sin buscar el máximo reconocimiento. Al igual que con la base de datos en alemán, se utilizaron audios de diferentes personas para obtener un sistema independiente del locutor. Al ser una base con sólo tres emociones, los resultados de las pruebas que se muestran a continuación son: variando la cantidad de componentes gaussianas en un modelo GMM y variando la cantidad de estados y la cantidad de gaussianas en cada estado en un modelo HMM. En la parametrización de los audios en español, se incorporó el coeficiente de energía, pues mejoró el rendimiento. También aquí, todos los resultados corresponden a porcentajes globales obtenidos de la validación cruzada de las 10 particiones de audios elegidos de forma aleatoria. Al igual que con la base de datos en alemán, el objetivo de estos experimentos fue determinar cual de los dos métodos de clasificación funciona mejor.

5.2.1. Reconocimiento de emociones con GMM

La Figura 5.3 muestra la tasa de reconocimiento obtenida al realizar un experimento con modelo GMM para diferentes cantidades de componentes gaussianas. Se puede apreciar que el reconocimiento del sistema varía de acuerdo a la cantidad de componentes gaussianas.

Los valores obtenidos para un clasificador GMM alcanzaron el mayor porcentaje acierto con 18 componentes gaussianas. La matriz de confusión de la Tabla 5.11 muestra que la mayor causa de errores fueron en sustituciones del estado neutral por el resto de las emociones. Esto es, la alegría fue reconocida con un 55 % de aciertos, el enojo con un 81 % y el estado neutral con un 44 %.

5.2.2. Reconocimiento de emociones con HMM

Los resultados que se muestran a continuación corresponden a los obtenidos al variar la cantidad de estados y la cantidad de componentes gaussianas en la mezcla de cada estado en un modelo HMM.

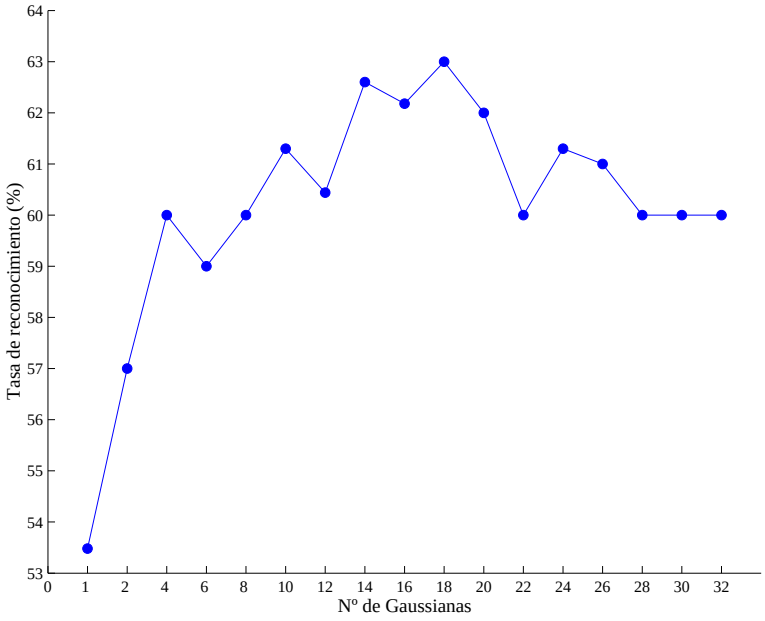

Figura 5.3. Reconocimiento de emociones en función de la cantidad de gaussianas con GMM [BD en español]

En la Figura 5.4 se puede apreciar la tasa de reconocimiento, promedio sobre las diez particiones, al variar la cantidad de estados involucrados en un modelo HMM. El promedio de cada barra vertical es obtenido manteniendo fija la cantidad de estados y variando la cantidad de componentes gaussianas en la mezcla. El reconocimiento del sistema no varía significativamente en función de la cantidad de estados, por lo que se presentan los resultados obtenidos para modelos HMM de 2 estados. En cada experimento se incrementa el número de estados HMM y se adiciona gradualmente la cantidad de componentes gaussianas en la mezcla hasta llegar la cantidad máxima deseada.

El experimento realizado en un modelo HMM con dos estados y 14 componentes gaussianas obtuvo una tasa de acierto del 60.86 %. En la matriz de confusión de la Tabla 5.12 se puede distinguir las sustituciones de una emoción por otra.

En el caso de un modelo con 3 estados, la mayor tasa de reconocimiento se alcanzó con 32 componentes gaussianas en los estados del modelo. Se obtuvo un 65.21 % de rendimiento y en la Tabla 5.13 se ve la matriz de confusión.

Al evaluar el rendimiento del reconocedor de emociones con 4 esta-

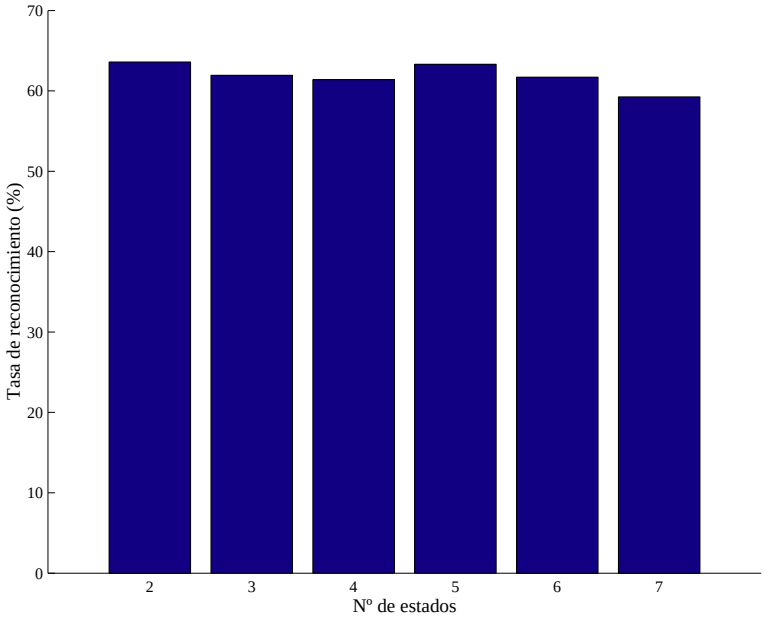

Figura 5.4. Reconocimiento de emociones en función de la cantidad de estados utilizando HMM [BD en español]

Tabla 5.12. Matriz de confusión para 2 estados con HMM (14 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**44**	9	17
Enojo	13	**54**	13
Neutral	24	14	**42**

Tabla 5.13. Matriz de confusión para 3 estados con HMM (32 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**36**	15	19
Enojo	7	**66**	7
Neutral	17	15	**48**

dos (Tabla 5.14), se obtuvo una tasa de reconocimiento del 61.74 % con 28 componentes gaussianas. En este caso, la tasa bajó considerablemente con respecto al experimento anterior.

En la Tabla 5.15 se puede apreciar los resultados obtenidos al realizar las pruebas con un reconocedor HMM con 5 estados y 4 componentes gaus-

Tabla 5.14. Matriz de confusión para 4 estados con HMM (28 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**33**	13	24
Enojo	10	**62**	8
Neutral	20	13	**47**

Tabla 5.15. Matriz de confusión para 5 estados con HMM (4 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**39**	9	22
Enojo	8	**61**	11
Neutral	21	15	**44**

Tabla 5.16. Matriz de confusión para 6 estados con HMM (6 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**51**	9	10
Enojo	12	**62**	6
Neutral	26	15	**39**

Tabla 5.17. Matriz de confusión para 7 estados con HMM (26 gaussianas)[BD en español]

Emoción	Alegría	Enojo	Neutral
Alegría	**38**	11	21
Enojo	9	**63**	8
Neutral	25	7	**48**

sianas en las mezclas de los estados. En este caso el porcentaje de acierto fue del 62.60 %.

Al incorporar un estado más al modelo HMM se obtuvo un 66.08 % de acierto con 6 componentes gaussianas en las mezclas de los estados. Este es el caso de mayor reconocimiento. Se puede apreciar la mejora de reconocimiento de la emoción alegría, mientras que el estado neutral disminuyó su tasa de reconocimiento. Puede distinguirse ésto en la Tabla 5.16.

El reconocedor de emociones con un modelo HMM de 7 estados no arrojó mejoras notorias. En la Tabla 5.17 se puede apreciar la matriz de confusión obtenida con 26 componentes de gaussianas, obteniendo un 64.78 % de reconocimiento.

A medida que se incrementa la cantidad de estados en el modelo, se

incorporan más variables y más parámetros deben ser estimados. Todo ésto afecta directamente al costo computacional (principalmente en la etapa de entrenamiento).

Se puede apreciar que la tasa de reconocimiento es menor que la obtenida con la base de datos en alemán. Son varios los factores que explican esta realidad. Una circunstancia a tener en cuenta es que el sistema de reconocimiento se ve influenciado por el canal de comunicaciones ya que los audios presentan niveles de ruido, distorsiones lineales e interferencias.

Analizando los resultados del test perceptual, en muchos casos los audios no son claros respecto de la emoción que expresan, dando lugar a subjetividades. Esto podría ser causante de malos resultados en el reconocedor de emociones. Se puede apreciar el caso de mayor subjetividad con el tipo de emoción *neutral*, donde el test perceptual alcanzó un 90.2 % de reconocimiento pero la tasa de reconocimiento en un modelo con 3 estados fue del 60 %. La emoción *enojo* obtuvo un 97.7 % de reconocimiento en el test perceptual, mientras que el reconocedor de emociones con 3 estados alcanzó una tasa de reconocimiento del 82.5 %. En el caso de la *alegría*, el test perceptual logró un 79 % y el reconocedor de emociones con 6 estados alcanzó un 72.85 % de reconocimiento.

Capítulo 6

Conclusiones y trabajos futuros

Los objetivos propuestos para este proyecto final de carrera se han cumplido satisfactoriamente y el mismo ha sido fuente de una publicación en congreso internacional [41]. En este trabajo se estudiaron dos modelos estadísticos para el reconocimiento de emociones, uno estático (GMM) y uno dinámico (HMM). Se han utilizado dos bases de datos para probar la eficiencia del reconocedor en dos idiomas distintos. Con la base de datos en alemán se evaluaron de tres hasta siete emociones distintas y con la base de datos en español, tres emociones. Se realizaron pruebas variando el número de componentes en los GMM y con diversa cantidad de estados y componentes gaussianas en los estados para el caso de HMM.

La base de datos en español fue desarrollada para evaluar el reconocedor en nuestro idioma, y debido a que los audios poseen baja calidad, nos limitaremos a concluir que los resultados son aceptables pero no suficiente a nivel de una aplicación real, puesto que sería necesario un modelo más robusto para este tipo de audios, donde se presenta mucho ruido.

Los resultados, sobre la base de datos en alemán, mostraron que las características espectrales desempeñan un papel significativo en el reconocimiento de emociones. Debido a que la forma del tracto vocal puede cambiar en virtud de los diferentes estados de ánimo, las características espectrales de discurso difieren para diversas emociones, incluso cuando se pronuncia una misma frase. Se obtuvieron resultados equivalentes a publicaciones recientes con características prosódicas.

A diferencia de trabajos anteriores, los resultados obtenidos pueden generalizarse respecto del hablante dado que el corpus de emociones uti-

lizado contempla elocuciones de diez hablantes distintos (cinco mujeres y cinco hombres). Las mezclas de gaussianas presentan un buen rendimiento que decae en función de la cantidad de emociones que se modelen, y es allí donde los HMM mejoran el rendimiento puesto que permiten modelos más complejos obteniendo un sistema con mayor tasa de aciertos.

En el presente proyecto final de carrera se logró integrar los conocimientos adquiridos en los años de carrera con los conocimientos devenidos de las tareas de investigación realizadas en el Grupo de investigación en señales e **in**teligencia **c**omputacional (**sinc**(i)), de esta facultad.

La prosodia y las características espectrales desempeñan un papel importante en la tarea de reconocimiento de la emoción. Por lo tanto, como trabajo futuro, es importante encontrar la manera de hacer una combinación de estas características para aumentar el rendimiento en sistemas de reconocimiento de emociones.

Si bien el sistema puede funcionar como un módulo de reconocimiento, aún queda pendiente la realización de una interfaz gráfica amigable al usuario a fin de extender el uso del reconocedor a distintos ambientes y dispositivos.

Apéndice A

Corpus en alemán

Frases expresadas por los actores:

- Der Lappen liegt auf dem Eisschrank. (Los trapos yacen sobre el refrigerador.)
- Das will sie am Mittwoch abgeben. (Ella lo entregará el Miércoles.)
- Heute Abend könnte ich es ihm sagen. (Esta noche yo podría comentarle eso.)
- Das schwarze Stück Papier befindet sich da oben neben dem Holzstück. (Ese fragmento negro de papel se encuentra allá arriba al lado del trozo de madera.)
- In sieben Stunden wird es soweit sein. (En siete horas estará.)
- Was sind denn das für Tüten, die da unter dem Tisch stehen? (Qué son esas bolsas, ellas están allí bajo la mesa?)
- Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter. (Ellos lo han subido y ahora ellos van hacia abajo otra vez.)
- An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. (Los fines de semana siempre iba a casa y recibi la visita de Agnes.)
- Ich will das eben wegbringen und dann mit Karl was trinken gehen. (Yo igualmente desviaria eso y entonces fuimos de copas con Karl.)
- Die wird auf dem Platz sein, wo wir sie immer hinlegen. (Aquello estará en el lugar, donde nosotros siempre los colocamos.)

Apéndice B

Corpus en español

Películas utilizadas para la extracción de los audios:

Motivos para no enamorarse. Argentina (2008). Intérprete seleccionado: *Jorge Marrale*.

Las manos. Argentina (2006). Intérprete seleccionado: *Jorge Marrale*.

El día que me amen. Argentina (2003). Intérprete seleccionado: *Jorge Marrale*.

Cenizas del paraíso. Argentina (1997). Intérpretes seleccionados: *Cecilia Roth y Jorge Marrale*.

Soy tu aventura. Argentina (2003). Intérprete seleccionado: *Jorge Marrale*.

La señal. Argentina - España (2007). Intérprete seleccionado: *Diego Peretti*.

Tiempo de valientes. Argentina (2005). Intérprete seleccionado: *Diego Peretti*.

Quien dice que es fácil. Argentina (2006). Intérpretes seleccionados: *Diego Peretti y Carolina Peleritti*.

No sos vos, soy yo. Argentina (2004). Intérprete seleccionado: *Diego Peretti*.

El nido vacío. Argentina (2008). Intérprete seleccionado: *Cecilia Roth*.

Deseo. Argentina - España (2002). Intérprete seleccionado: *Cecilia Roth.*

Kamchatka. Argentina (2002). Intérprete seleccionado: *Cecilia Roth.*

Una noche con Sabrina love. Argentina (2000). Intérprete seleccionado: *Cecilia Roth.*

Todo sobre mi madre. España (1999). Intérprete seleccionado: *Cecilia Roth.*

Taxi un encuentro. Argentina (2000). Intérprete seleccionado: *Diego Peretti.*

XXY. Argentina (2007). Intérprete seleccionado: *Carolina Peleritti.*

Antigua vida mía. Argentina (2001). Intérpretes seleccionados: *Cecilia Roth y Jorge Marrale.*

Historia de sexo de gente común. Argentina (serie, 2004-2005). Intérprete seleccionado: *Carolina Peleritti.*

Vidas Robadas. Argentina (novela, 2008). Intérprete seleccionado: *Jorge Marrale.*

Simuladores. Argentina (serie, 2003). Intérprete seleccionado: *Diego Peretti.*

Bibliografía

[1] Planet, S., Morán, J.A., Formiga, L.: Reconocimiento de emociones basado en el análisis de la señal de voz parametrizada. In Actas da 1a Conferência Ibérica de Sistemas e Tecnologias de Informação, Ofir, Portugal **2** (2006) 837–854

[2] Cowie, R., et al., N.T.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18**(1) (2001) 32–80

[3] Goleman, D.: Inteligencia emocional. Editorial Kairos. Barcelona (1996)

[4] Izard, C.: Organizational and motivational functions of discrete emotions. En M. Lewis, J.M. Haviland (Eds), Guilford Press, New York (1993)

[5] Picard, R.W.: Affective Computing. MIT Press (1997)

[6] Darwin, C.: The expression of the emotions in man and animals. John Murray (1872)

[7] Cornelius, R.: The Science of Emotion Research and Tradition in the Psychology of Emotion. Prentice Hall (1996)

[8] Scherer, K.R.: Personality markers in speech. Cambridge University Press, Cambridge (1979)

[9] RAE: Diccionario de la Real Academia Española, 22a. edición., Espasa Calpe: Madrid. (2001)

[10] Causa, E., Sosa, A.: La computación afectiva y el arte interactivo. In Projecto BIOPUS (2007)

[11] Arana, A.: Escuchando la voz de las emociones, http://www.degerencia.com/articulo/escuchando_la_voz_de_las_emociones. (2007)

[12] Mehrabian, A.: Communication without words. Psychology Today **2** (1968) 53–56

[13] García, J.C.P.: Interfaces Afectivas Síncronas. PhD thesis, Universidad de las Américas Puebla, Cholula, Puebla, México (2004)

[14] Cowie, R., Cornelius, R.: Describing the emotional states that are expressed in speech. Speech Communication **40**(1) (2003) 5–32

[15] Lin, Y.L., Wei, G.: Speech emotion recognition based on HMM and SVM. Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on **8** (18-21 August 2005) 4898–4901 Vol. 8

[16] Gil, L., et al.: Reconocimiento automático de emociones utilizando parámetros prosódicos. Procesamiento del lenguaje natural (35) (September 2005) 13–20

[17] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2 sub edn. Wiley-Interscience (October 2000)

[18] Bishop, C.M.: Pattern Recognition and Machine Learning. 1 edn. Springer (2006)

[19] Noguerias, A., Moreno, A., Bonafonte, A., Mariño, J.: Speech Emotion Recognition Using Hidden Markov Models. Eurospeech 2001 (2001) 2679–2682

[20] Thomas, D.L., Diener, E.: Memory accuracy in the recall of emotions. Journal of Personality and Social Psychology **59** (1990) 291–297

[21] Staats, A.: A psychological behaviorism theory of personality. En T. Millon and M.J. Lerner (Eds.) Handbook of Psychology (2003)

[22] Ekman, P., Friesen, W.V.: Facial Action Coding System Investigator's Guide. Consulting Psycologist Press, Palo Alto, CA (1978)

[23] Oppenheim, A.V., Wilsky, A.S.: Señales y Sistemas. Prentice Hall (1998)

[24] Milone, D.H.: Información Acentual para el Reconocimiento Automático del Habla. PhD thesis, Universidad de Granada, Granada, España (Marzo 2003) Memoria de Tesis.

[25] Kuc, R.: Introduction to digital signal processing. McGraw-Hill Book Company (1988)

[26] Deller, J.R., Proakis, J.G., Hansen, J.H.: Discrete-Time Processing of Speech Signals. Macmillan Publishing, New York (1993)

[27] Albornoz, E.M.: Sistema de análisis prosódico y reconocimiento automático del habla. Proyecto final de carrera, ingeniería en informática, Universidad Nacional del Litoral (2006)

[28] Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)

[29] Milone, D.H.: Fundamentos del reconocimiento automático del habla. Technical report, Universidad Nacional del Litoral (2004)

[30] Rabiner, L.R., Juang, B.H.: An introduction to hidden Markov models. IEEE Acoustics Speech and Signal Processing Magazine **3**(1) (January 1986) 4–16

[31] Albornoz, E.M.: Modelado de estructuras acentuales a partir de rasgos prosódicos básicos con modelado ocultos de markov y su incorporación a un sistema de reconocimiento automático del habla. Technical report, Universidad Nacional del Litoral (2005)

[32] Baum, L., Egon, J.: An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology. Bull. Amer. Meteorol. Soc. **73** (1967) 360–363

[33] Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. Ann Math **37** (1966) 1554–1563

[34] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. Proc. Interspeech 2005 (September 2005) 1517–1520

[35] Douglas-Cowie, E., Cowie, R., Schröder, M.: A new emotion database: Considerations, sources and scope. In: Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland (September 2000) 5–7

[36] Paeschke, A.: Global Trend of Fundamental Frequency in Emotional Speech. In: ISCA - Speech Prosody, Nara, Japan (March 2004) 671–674

[37] Paeschke, A., Sendlmeier, W.F.: Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In: ISCA - ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK (September 2000) 75–80

[38] Burkhardt, F., Sendlmeier, W.F.: Verification of Acoustical Correlates of Emotional Speech Using Formant-Synthesis. In: ISCA - Speech Prosody, Newcastle, Northern Ireland, UK (September 2000) 151–156

[39] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.1). Cambridge University Engineering Department., Cambridge, Inglaterra. (December 2001)

[40] Michie, D., Spiegelhalter, D., Taylor, C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood, University College, London (1994)

[41] Albornoz, E., Crolla, M.B., Milone, D.: Recognition of emotions in speech. En los anales de CLEI 2008 (Septiembre 2008) 1120–1129