PROYECTO FIN DE CARRERA

AUTORA: Carmen Rincón Llorente

TUTOR: D. Roberto Barra Chicote

PONENTE: D. Juan Manuel Montero Martínez

TÍTULO: Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz.

DEPARTAMENTO: Ingeniería Electrónica de la Escuela Técnica Superior de Ingenieros de Telecomunicación de Madrid (Universidad Politécnica de Madrid).

GRUPO: Tecnología del Habla.

TRIBUNAL

El tribunal nombrado para evaluar dicho proyecto está compuesto por los docentes:

Presidente: D. Javier Macias Guarasa

Vocal: D. Rubén San Segundo Hernández

Secretario: D. Juan Manuel Montero Martínez

Suplente: D. Fernando Fernández Martínez

CALIFICACIÓN:

FECHA DE LECTURA Y DEFENSA: Abril de 2007

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS
DE TELECOMUNICACIÓN

PROYECTO FIN DE CARRERA

Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz

AUTORA: Carmen Rincón Llorente
TUTOR: Roberto Barra Chicote
PONENTE: Juan Manuel Montero Martínez

FECHA: Abril de 2007

RESUMEN

Uno de los principales retos de la Tecnología del Habla es desarrollar sistemas de interacción con el usuario lo más humanos posibles, para los que es necesario un conocimiento de las emociones del usuario. En este proyecto se lleva a cabo el diseño y desarrollo de un sistema automático de identificación de emociones a través de características extraídas de la voz, relacionadas con rasgos segmentales y con rasgos prosódicos. El sistema está formado por cuatro bloques: el *parametrizador*, que nos permite extraer los parámetros propios de la voz; el bloque de *normalización*, mediante el cual intentamos reducir la variabilidad de las características de la voz; el bloque de *entrenamiento*, con el que extraeremos unos modelos propios para cada emoción a partir de los vectores con características de la voz; y el *clasificador*, que nos permitirá decidir con que emoción se corresponde una determinada señal de voz de entrada, basándose en los modelos obtenidos en el entrenamiento. Para evaluar el sistema desarrollado se dispondrá de dos bases de datos en distinto idioma: una castellana y otra alemana.

Palabras clave: identificación de emociones, GMM, MFCC, análisis prosódico, normalización cepstral.

AGRADECIMIENTOS

Muchas personas son a las que tengo agradecer tanto el haber sacado adelante este proyecto como la carrera en general.

Los primeros de todos creo que deben ser mis padres, por darme siempre todo lo que he necesitado, por animarme en todo momento, por darme todo el cariño del mundo y por estar siempre a mi lado.

Otra persona muy importante a la que debo agradecer haber sacado la carrera tan bien es a Eloy, por enseñarme a tener mucha confianza en mi misma, ayudarme siempre que he tenido malos momentos, aconsejarme en cualquier tema, ayudarme con mis problemas técnicos y por seguir estando a mi lado, a pesar de todo.

Pensando en este proyecto, a la persona que más tengo que agradecer es a Roberto, por su paciencia, por sus consejos, por enseñarme tantas cosas nuevas, por sacar siempre un ratillo para ayudarme, por poner tanto entusiasmo en todas las cosas, por recuperar mi ordenador cuando se estropeo, por poner tanta confianza en mi,...

Gracias a Evita, mi compañera de clase, gracias a la que he sacado adelante tantas asignaturas, por esas horas en el laboratorio de programación, por esas comidas en cafetería, por sus consejos y por todo, muchas gracias, guapa. Espero recuperar la amistad que últimamente hemos perdido.

Gracias a Pati, mi compañera (gaditana, segoviana y parisina) de laboratorios, por las horas pasadas en el laboratorio para conseguir que funcionase la placa de LCEL (incluidas las horas pasadas en mi casa, con el laboratorio montado en la cocina). Y a Gión por traernos esos kit-kat para recuperar energía y poder seguir con nuestra placa.

Muchas gracias también a Vane, mi compañera inseparable de proyecto, por estar siempre ahí para escucharme, darme muy buenos consejos, por esas comidas en Ericsson, esas largas conversaciones y por enseñarme a ver el mundo de otra manera muy diferente a la que yo le veía.

Gracias a mis compañeros de la carrera: Gela, Sara, Merche, Lara, Ari, Ana Herrero, Ana Do Carmo, Blanca, Belén, Alicia, Guille, Willy, Carlos, Hugo, Pablo, Juan, Diego, Galán, Víctor, Iván, Pas, Jaime, María Nogales, Viveca, Alex,…, por todas las fiestas, los momentos en cafetería, los descansos de clase y tantas cosas. Muy especialmente gracias a Guada, Mery y Helen por los momentos inolvidables vividos

sobre todo en el último año, los viajes y las fiestas.

Gracias a Miguel por ser tan buen compañero de clase, saber que puedo contar con él para lo que sea y ayudarme siempre que lo he necesitado.

Gracias a mi hermano por intentarme enseñar siempre a ser un poco autodidacta, darme siempre buenos consejos y por prestarme su habitación y su ordenador para acabar sin problemas el proyecto. Gracias a mi hermana por hacerme siempre todos los favores que he necesitado.

Gracias a mis amigos de siempre. A Isa por cuidarme tanto, por esas cenas tan ricas, por esas noches de fiesta, por esos oídos tan grandes: no cambies nunca. A Elena por ese cariño que siempre irradia, por estar siempre muy atenta de mí, por los viajes hechos y los que nos quedan por hacer. A Almu, a Rakel, a Mariu y a Alfonso, porque se que siempre estaréis ahí para lo que necesite. A Guille, Tru y Sergi, por esas partidas de dardos, por las horas pasadas en San José y por los buenos momentos de los veranos.

Gracias a Esther por esas canciones tan bonitas que me han hecho pasar el tiempo más ameno mientras las escuchaba.

Gracias a Nico por aparecer en el momento en el que más le necesitaba, por ser un solete, por esas largas conversaciones que me han ayudado a llevar mejor el proyecto, por hacer tantas veces de almohada para llorar y por sus abrazos.

Gracias a Juancho por elegirme para realizar este proyecto, por darme un curso personalizado de C y por estar siempre atento sobre como va mi proyecto. Gracias a Javier Macías por animarme a coger este proyecto cuando no estaba muy convencida de ello y por todos sus consejos.

Gracias a mis compañeros del laboratorio: Rosi, Rosalía, Isa, Amparo, Jose, Fran, Vicente,... También gracias a Luisfer. Y a Julián.

Gracias a María, Laura y Gonzalo, mis compañeros del master. También gracias a Bea, por darme ese gran abrazo en un momento de bajón, que me ayudo bastante para seguir adelante.

Gracias a Nacho, Pablo, José Luis, Borja y Gonzalo, mis compañeros de Ericsson.

Gracias a Dani por preocuparse tanto por mí y perdonarme que últimamente le tenga un poco olvidado.

Gracias a todos los que haya podido olvidar, pero que llevo en mi corazón.

ÍNDICE

ÍNDICE DE TABLAS Y FIGURAS

- ***Tablas***

Figuras

LISTADO DE SIGLAS Y ABREVIATURAS

- ***Siglas:***

CMN: Cepstral Mean Normalization.

CVN: Cepstral Varianze Normalization.

DES: Danish Emotional Speech database.

DCT: Discrete Cosine Transform.

EMODB: Berlin Database of Emotional Speech.

FFT: Fast Fourier Transform.

FT: Fourier Transform.

GMM: Gaussian Mixture Model.

HMM: Hidden Markov Models.

LDA: Linear Discriminant Analysis.

MFCC: Mel Frequency Cepstral Coefficients.

PCA: Principal Component Analysis.

SES: Spanish Emotional Speech database.

SEV: Spanish Emotional Voices.

- ***Abreviaturas:***

Hz.: hercios.

dB.: decibelios.

1. INTRODUCCIÓN

El concepto de emoción es difícil de definir, pero combinando múltiples definiciones podemos decir que se trata de un estado mental, consciente o no, de una cierta intensidad y duración más o menos breve, que puede actuar como favorecedor u obstaculizador de las reacciones humanas ante determinados eventos externos o internos, y que puede provocar ciertas alteraciones fisiológicas prototípicas, perceptibles desde el exterior. Las emociones las podemos dividir en primarias (alegría, enfado, sorpresa, miedo, asco y tristeza) y secundarias (derivadas de las primarias y combinaciones de varias de ellas).

El estado emocional del individuo hace variar las características de la voz respecto a su estado neutro (ausencia de emoción) 1.1.1.45; viéndose afectadas tanto las características segmentales, como son la frecuencia, la amplitud y la fase; como los rasgos prosódicos, como son el tono fundamental, el ritmo y la energía. Para poder modelar a través de la voz el estado emocional del individuo, debemos conocer en qué medida influye cada una de las emociones en los parámetros de las características citadas.

La comunicación verbal tiene una importancia especial ya que, al menos las emociones primarias, tienen un patrón específico y universal para su comunicación, configurado por la postura corporal, la expresión facial, y la prosodia del lenguaje, es decir, el tono emocional del habla 1.1.1.45.

Las emociones suelen tener una función comunicativa y valorativa, por lo que su reconocimiento por medio de la voz puede ayudar a mejorar el gestor de diálogo de los sistemas de interacción persona-máquina. También ayudará a mejorar la naturalidad del “Feed-Back” que recibe el usuario, sustituyendo la voz sintética estándar, que suele responder a un modelo de voz neutra sin ningún matiz emocional.

En el desarrollo de este proyecto, nos hemos centrado exclusivamente en la tarea

de identificación de emociones, completando otros proyectos del Grupo de Tecnología del Habla que abordan la labor de síntesis de emociones.

En resumen, un sistema de identificación de emociones consta principalmente de tres fases: la primera de ellas sería la de *parametrización*, en la cual extraeremos diversos parámetros de las características de la voz; la segunda sería el *entrenamiento*, en la que generaremos un modelo para cada una de las emociones, lo más preciso posible, a partir de los parámetros extraídos en la fase anterior; y la ultima fase sería la de *clasificación*, en la que a partir de los modelos generados en la fase anterior podamos decidir a qué tipo de emoción pertenece un ejemplo de voz pasado como entrada. Adicionalmente, nuestro sistema dispondrá de una fase de *normalización* de los vectores que contengan las características de la voz, con el objetivo de intentar reducir la variabilidad presente en ésta.

El objetivo prioritario de este proyecto no es evaluar los posibles algoritmos empleados en la clasificación, sino, basándose en el algoritmo de mezcla de gausianas (GMM – Gaussian Mixture Model), estudiado en 1.1.1.45, 1.1.1.45 y 1.1.1.45, evaluar el impacto de las características de la voz en la tarea de identificación de emociones.

El contenido de la presente memoria del proyecto realizado sigue la siguiente estructura: como punto inicial se presenta una descripción de los aspectos básicos de las emociones, realizando una descripción detallada de cada una de las emociones consideradas en el proyecto (sorpresa, alegría, tristeza, enfado, asco, miedo y aburrimiento). A continuación se describen las bases de datos utilizadas para evaluar el sistema desarrollado, así como las características particulares de las emociones interpretadas por los actores en dichas bases de datos. El siguiente capítulo aborda la descripción del sistema desarrollado para el reconocimiento automático de emociones a través de características extraídas de la voz, con una descripción detallada de cada una de las fases en las que podemos dividirlo (parametrización, normalización, entrenamiento y clasificación). Una vez diseñado y desarrollado el sistema, es hora de pasar a evaluarlo, para ello se realizarán una serie de experimentos que dividiremos en dos tipos fundamentales: aquellos basados en características segmentales y los basados en características prosódicas. Dentro de los primeros, podemos hacer otra división: experimentos con la base de datos castellana, con la alemana y con ambas, de forma que abordemos la tarea de identificación de emociones de distintos idiomas. En identificación

de emociones basada en rasgos prosódicos sólo trabajamos con la base de datos castellana. Por tanto, en los capítulos 6 y 7 se realiza una descripción detallada de cada uno de los experimentos realizado, así como un análisis de los resultados obtenidos y se extraen conclusiones particulares de dichos resultados. Finalmente, se exponen las principales conclusiones a las que llegamos una vez analizados todos los experimentos realizados y se proponen una serie de líneas futuras de trabajo que se abren como consecuencia de este proyecto.

2. OBJETIVOS

Los principales objetivos que nos propusimos abordar al inicio de este proyecto fueron los siguientes:

- Modificación y adaptación a un entorno de aplicación Windows de las rutinas de identificación basadas en rasgos segmentales disponibles.

- Análisis de las similitudes y diferencias sobre la naturaleza de las emociones en la voz, entre los idiomas de las dos bases de datos disponibles en el departamento (castellano y alemán).

- Aplicación de técnicas de normalización a las características extraídas de la voz, intentando evitar su variabilidad, y evaluación de la posible mejora obtenida al emplear las características, previamente normalizadas, en el sistema desarrollado. Las técnicas de normalización que aplicaremos serán las basadas en la media y/o en la varianza estimadas respecto a la voz del locutor o respecto a la voz neutra de dicho locutor.

- Diseño e implementación de técnicas de parametrización basadas en aspectos prosódicos, extraídos a partir de los grupos fónicos en los que podemos segmentar la locución. Estas características pueden estar relacionados con el contorno de F0 (valor medio, máximo, mínimo y rango de F0, así como pendiente de subida y de bajada); o con el ritmo (velocidad de locución de la frase y velocidad de locución de los grupos fónicos).

- Análisis de la calidad de las distintas características prosódicas para la identificación de emociones.

- Evaluación de las estrategias implementadas, centrándose principalmente en:

    - La evaluación de los diversos parámetros extraídos de las características

de la voz, según la naturaleza dominante de cada emoción.

  - La interpretación de los resultados a la vista de su fiabilidad estadística, considerando las mejoras y degradaciones observadas.

- Comparación de los resultados obtenidos mediante el reconocimiento automático de emociones con las conclusiones obtenidas en estudios perceptuales previos.

3. ASPECTOS BÁSICOS DE LAS EMOCIONES

A lo largo de este capítulo intentaremos desarrollar las características más destacadas del proceso emocional. Comenzaremos con una definición del concepto de emoción, para explicar a continuación las principales funciones que tienen cada una de las emociones primarias: función adaptativa, social y motivacional. Describiremos también en qué consiste la activación emocional, que incluye una experiencia subjetiva, una expresión corporal, el afrontamiento y el soporte fisiológico necesario para la ejecución de estas respuestas. Finalmente se describirá en detalle la activación asociada a cada una de las emociones consideradas en este proyecto y se hará una comparación entre sus principales características prosódicas.

3.1. Concepto de emoción

Definir qué es una emoción es una tarea muy complicada. Los autores Kleinginna y Kleinginna 1.1.1.45 recopilaron más de cien definiciones de emoción, pudiendo agrupar las diferentes formas de conceptualizar la emoción en once categorías (afectiva, cognitiva, basada en estímulos elicitadores, fisiológica, emocional/expresiva, disruptiva, adaptativa, multifactorial, restrictiva, motivacional y escéptica). Estos autores llegaron a la siguiente definición de emoción:

"Un complejo conjunto de interacciones entre factores subjetivos y objetivos, mediadas por sistemas neuronales y hormonales que: (a) pueden dar lugar a experiencias efectivas como sentimientos de activación, agrado-desagrado; (b) generar procesos cognitivos tales como efectos perceptuales relevantes, valoraciones, y procesos de etiquetado; (c) generar ajustes fisiológicos…; y (d) dar lugar a una conducta que es frecuentemente, pero no siempre, expresiva, dirigida hacia una meta y adaptativa" (Kleinginna y Kleinginna 1.1.1.45).

Según esta definición, las emociones tienen un carácter multidimensional y podemos entenderlas como un proceso que implica una serie de condiciones desencadenantes (estímulos relevantes), la existencia de experiencias subjetivas o sentimientos (interpretación subjetiva), diversos niveles de procesamiento cognitivo (procesos valorativos), cambios fisiológicos (activación), patrones expresivos y de comunicación (expresión emocional), que tiene unos efectos motivadores (movilización para la acción) y una finalidad: que es la adaptación a un entorno en continuo cambio.

3.2.Función de las emociones

Las funciones que todas las emociones deben cumplir, que las hace útiles y beneficiosas son:

- **Función adaptativa**: prepara al organismo para que ejecute eficazmente una conducta exigida por las condiciones ambientales, que movilice la energía necesaria para ello y que dirija la conducta a un objetivo determinado.

  Las funciones adaptativas de las emociones primarias son:

  - Sorpresa: exploración.
  - Asco: rechazo.
  - Alegría: afiliación.
  - Miedo: protección.
  - Ira: autodefensa.
  - Tristeza: reintegración.

- **Función social:** comunica nuestro estado de ánimo.

Se basa en la expresión de las emociones, lo cual permite a las demás personas predecir el comportamiento que vamos a desarrollar y a nosotros el suyo, lo que tiene un indudable valor en los procesos de relación interpersonal. De esta forma, la expresión de las emociones puede considerarse como una serie de estímulos discriminativos que facilitan la realización de conductas sociales.

Estas funciones se cumplen mediante varios sistemas de comunicación diferentes: la comunicación verbal (o información a los demás de nuestros sentimientos), la comunicación artística y la comunicación no verbal. La comunicación verbal tiene una importancia especial ya que, al menos las emociones primarias, tienen un patrón específico y universal para su comunicación, configurado por la postura corporal, la expresión facial, y la prosodia del lenguaje, es decir, el tono emocional del habla.

Esta función facilita la interacción social, influye en la conducta de los demás, permite la comunicación de los estados afectivos y promueve la conducta prosocial.

- **Función motivacional:** facilita las conductas motivadas.

Una emoción puede determinar la aparición de la propia conducta motivada, dirigirla hacia determinada meta y hacer que se ejecute con un cierto grado de intensidad.

3.3.Activación emocional

La activación emocional es de carácter multifactorial e implica múltiples efectos, entre los cuales los más específicos son: una experiencia o efecto subjetivo, una expresión corporal o efecto social, un afrontamiento o efecto funcional y un soporte fisiológico para la ejecución de todas las respuestas anteriores. A continuación explicamos brevemente cada uno de estos efectos:

- Experiencia subjetiva: se refiere a las sensaciones o sentimientos que produce la respuesta emocional, cuya principal temática es de placer o displacer que se desprende de la situación, seguida por la de activación o intensidad.

- Expresión corporal: se refiere a la comunicación y exteriorización de las emociones mediante la expresión facial y otra serie de procesos de comunicación no verbal tales como los cambios posturales o la entonación. Además, la expresión emocional cumple otras funciones como la de controlar la conducta del receptor, ya que permite a éste anticipar las reacciones emocionales y adecuar su comportamiento a tal situación.

  Junto con la expresión facial, la prosodia o tono emocional del habla son las comunicaciones no verbales las que más informan sobre el estado emocional de

una persona. Globalmente se aprecia una relación entre el ritmo y la valencia afectiva, de tal forma que las emociones positivas son expresadas con un ritmo más regular que las emociones negativas.

- Afrontamiento: se refiere a los cambios comportamentales que producen las emociones y que hacen que las personas se preparen para la acción, es decir, al conjunto de esfuerzos cognitivos y conductuales, que están en un constante cambio para adaptarse a las condiciones desencadenantes.

- Soporte fisiológico: se refiere a los cambios y alteraciones que se producen en el sistema nervioso central, periférico y endocrino.

Todos estos elementos que configuran la activación emocional, serán vistos con más detalle en el siguiente apartado, en el que se describen las diferentes emociones.

Resumiendo, la función de la activación emocional es dirigir las actividades y las interacciones que rigen la percepción, la atención, el aprendizaje, la memoria, la elección de metas, las prioridades motivacionales, las estructuras categoriales y conceptuales, las reacciones fisiológicas, los reflejos, las reglas de decisión de comportamiento, el sistema motor, los procesos de comunicación, la determinación del nivel de energía y de esfuerzo, la coloración afectiva de los acontecimientos y estímulos, la valoración de las situaciones, etc.

3.4.Descripción de las emociones

Según Izard 1.1.1.45, los requisitos que debe cumplir cualquier emoción para ser considerada como básica o primaria son: tener un sustrato neural específico y distintivo, tener una expresión o configuración facial específica y distintiva, poseer sentimientos específicos y distintivos, derivar de procesos biológicos evolutivos, y manifestar propiedades motivacionales y organizativas de funciones adaptativas.

A continuación describimos, para cada una de las emociones consideradas en el proyecto, los principales desencadenantes que la provocan, así como la activación asociada a cada una de ellas, considerando los efectos subjetivos, el perfil psicofisiológico y la expresión vocal.

3.4.1.La emoción de sorpresa

La sorpresa es considerada como la emoción básica más singular. Se trata de una reacción emocional neutra, que no podemos clasificar como negativa o positiva, como agradable o desagradable. Sin embargo, la neutralidad es difícil de conseguir, así podemos tener sorpresas agradables y desagradables, pero escasas veces indiferentes. Algunas investigaciones indican que la sorpresa tiene más en común con las emociones negativas que con las positivas (de hecho, la expresión facial de la sorpresa se confunde más a menudo con emociones negativas, especialmente miedo, que con emociones positivas).

La sorpresa se produce por lo inesperado o desconocido. Puede ser descrita como una sensación causada por algún acontecimiento repentino e inesperado. Se produce de forma súbita ante una situación novedosa o extraña y desaparece con la misma rapidez con que apareció. Se trata de la emoción más breve de todas las primarias.

Hay autores que consideran esta emoción como primaria o básica, y otros que la consideran secundaria, derivada a partir de emociones primarias como la ira o el miedo. Izard la considera como una emoción básica, pero reconoce que no tiene todas las características propias de estas emociones.

1.1.1.1.Desencadenantes de esta emoción:

- Estímulos novedosos, de una intensidad entre débil y moderada.
- Aparición de acontecimientos inesperados o fuera de contexto.
- Aumentos bruscos de la intensidad en la estimulación.
- Interrupción inesperada o corte de una actividad en curso.

1.1.1.2.Activación asociada a la sorpresa

- Efectos subjetivos:
    - Mente en blanco.
    - Sensaciones de incertidumbre.

- Perfil psicofisiológico:
    - Actividad cardiovascular:
        - Desaceleración fásica de la frecuencia cardíaca.
        - Vasoconstricción periférica.
        - Vasodilatación cefálica.
    - Actividad electrodérmica:
        - Aumento brusco y fásico de la conductancia de la piel.
    - Actividad respiratoria y de la musculatura esquelética:
        - Aumento fásico del tono muscular general.
        - Interrupción puntual de la respiración, caracterizada por un cambio en la frecuencia y amplitud de la respiración o por una inspiración breve y de corta latencia.
        - Alta amplitud respiratoria.

- Expresión vocal:

Las exclamaciones de sorpresa se distinguen de las vocalizaciones de otras emociones tales como de la ira, miedo, vergüenza o placer. La emoción de sorpresa está fuertemente asociada con un tono de alto nivel. Las vocalizaciones espontáneas de la sorpresa son del tipo de ¡o!, ¡cómo!.

Los efectos de la sorpresa sobre la frecuencia fundamental y la prosodia, en comparación con el estado neutro, son los siguientes:

- Frecuencia fundamental y prosodia:
    - Tono medio mayor que el de la voz neutra.
    - Amplio rango de F0.
    - Velocidad de habla similar a la neutra.

3.4.2.La emoción de asco

El asco se considera una emoción básica porque tiene un sustrato neural innato, una expresión universal también innata, un único estado motivacional-afectivo y un patrón de respuesta asociado que es relativamente estable a lo largo de distintas situaciones, culturales e incluso especies.

El asco es una sensación que se refiere en primer lugar a algo que repugna al sentido del gusto, algo percibido en ese momento o imaginado con viveza, y en segundo lugar a algo que produce una sensación parecida en el sentido del olfato, de tacto, o incluso de la vista.

En el sentido más general el término asco define una marcada aversión producida por algo fuertemente desagradable o repugnante.

1.1.1.3.Desencadenantes de esta emoción

- Ciertos alimentos, comida putrefacta, maloliente.
- Secreciones corporales.
- Ciertos animales.
- Falta de higiene.

1.1.1.4.Activación asociada al asco

- Efectos subjetivos:
    - Repulsión.
    - Sentido de ofensa.
    - Percepción de que algo no es como debería ser.
- Perfil psicofisiológico:
    - Actividad gastrointestinal:
        - Si el estímulo es oloroso o gustativo, aparecen habitualmente sensaciones gastrointestinales desagradables, tales como las náuseas.

- Actividad cardiovascular:
  - Moderada elevación de la frecuencia cardíaca.
- Actividad electrodérmica:
  - Moderada elevación del nivel de la conductancia de la piel.
- Actividad respiratoria y de la musculatura esquelética:
  - Elevación en la tensión muscular general.
  - Elevación de la frecuencia respiratoria, con especial prolongación de las pausas entre inspiraciones.

- Expresión vocal:
  - Emisión de sonido como de aclarar la garganta.
  - Sonidos guturales *aj* o *uf*.

Los efectos del asco sobre la frecuencia fundamental y la prosodia, en comparación con el estado neutro, son los siguientes:

- Frecuencia fundamental y prosodia:
  - Tono medio bajo.
  - Amplio rango de F0.
  - Velocidad de locución más baja, con grandes pausas.

3.4.3.La emoción de miedo

El miedo es la emoción más estudiada tanto en los animales como en el hombre, ya que es una emoción que ha despertado mucho interés.

"El miedo es un estado emocional negativo o aversivo con una activación muy elevada que incita la evitación y el escape de las situaciones que amenazan la supervivencia o el bienestar del organismo" (Öhman, Flykt y Lundqvist 1.1.1.45).

"El miedo es una señal emocional de advertencia de que se aproxima un daño

físico o psicológico. El miedo también implica una inseguridad respecto de la propia capacidad para soportar o mantener una situación de amenaza. La intensidad de la respuesta emocional de miedo depende de la incertidumbre sobre los resultados" (Fernández-Abascal 1.1.1.45).

Hay que diferenciar la emoción de miedo de la ansiedad. Mientras que el miedo hace referencia a una emoción producida por un peligro presente e inminente, encontrándose ligado al estímulo que lo genera; la ansiedad hace referencia a la anticipación de un peligro futuro, indefinible e imprevisible, siendo la causa más vaga y menos comprensible que en el miedo.

1.1.1.5.Desencadenantes de esta emoción

- Percepción de daño o peligro.
- Estímulos muy intensos, muy notorios y nuevos.
- Peligros evolutivos especiales.
- Estímulos procedentes de interacciones sociales.
- Estímulos atemorizantes condicionados.

1.1.1.6.Activación asociada al miedo

- Efectos subjetivos:
    - Sensación de tensión o de gran activación.
    - Desasosiego.
    - Preocupación y recelo por la propia seguridad o por la salud.
    - Sensación de pérdida de control.

- Perfil psicofisiológico:
    - Actividad cardiovascular:
        - Aumento de la frecuencia cardíaca (las mayores de todas cuantas se producen en respuesta a una situación emocional).

- Aumento de la fuerza de contracción del corazón.
- Aumento de la presión arterial sistólica y diastólica.
- Aumento de la vasoconstricción periférica, cuya consecuencia es una apreciable disminución de la temperatura.

o Actividad electrodérmica:

- Aumento del nivel de conductancia de la piel y aumento en el número de fluctuaciones espontáneas de la misma.

o Actividad respiratoria y de la musculatura esquelética:

- Aumento de la tensión muscular.
- Aumento de la frecuencia respiratoria, acompañada de reducciones en su amplitud, es decir, se produce una respiración superficial e irregular.
- El miedo condicionado potencia el reflejo de sobresalto.

• Expresión vocal:

En situaciones de miedo extremo puede existir la tendencia natural a emitir gritos de alta frecuencia o chillidos. En particular, el miedo tiende a ser asociado tanto con la elevación como con la variabilidad del tono. Cuando una persona tiene miedo se suele producir un tartamudeo asociado al “jitter segmental” que se produce en la señal de habla.

Los efectos que produce el miedo sobre los diferentes parámetros vocales, en comparación con el estado neutro, son los siguientes:

o Fluencia:

- Mayor número de sílabas por segundo.
- Menor duración de la sílaba.
- Menor duración de las vocales acentuadas.

        - El número y duración de las pausas puede ser menor o mayor.

    - Frecuencia fundamental y prosodia:

        - Tono medio elevado.
        - Elevada desviación de F0.
        - Presenta el mayor rango de F0 de todas las emociones.
        - Gran número de cambios en la curva del tono.
        - Rápida velocidad de locución.

3.4.4.La emoción de alegría

La alegría es el sentimiento positivo, que surge cuando la persona experimenta una atenuación en su estado de malestar, cuando consigue alguna meta u objetivo deseado, o cuando tenemos una experiencia estética (por ejemplo, la visión de un rostro agraciado o la contemplación de una bella escultura).

Por lo general, la alegría es una experiencia emocional de duración breve, aunque, ocasionalmente, puede experimentarse como un estado de placer intenso.

1.1.1.7.Desencadenantes de esta emoción

- Atenuación de contingencias negativas.
- Acontecimientos positivos.
- Experiencia vivida por otros.
- Alegría hilarante, aquélla que cursa con sonrisas, risas o carcajadas: situaciones cómicas, estimulación táctil, transgresión de normas o tabúes, …

1.1.1.8.Activación asociada a la alegría

- Efectos subjetivos:
    - Vivencia placentera y de carácter reforzante.

- Actitud optimista.
- Aumento de la autoestima y de la autoconfianza.

- Perfil psicofisiológico:
  - Actividad cardiovascular:
    - El hecho de esbozar una sonrisa produce una ligera aceleración de la frecuencia cardíaca, que se hace más acusada cuando el desencadenante es la risa.
    - Aumento de los niveles de presión sanguínea sistólica y diastólica.
    - Aumento del volumen sanguíneo periférico.
  - Actividad electrodérmica:
    - Fluctuaciones que no se limitan al acto de la risa, sino que aparecen también cuando se inducen estados de alegría sin manifestaciones jocosas.
  - Actividad respiratoria y de la musculatura esquelética:
    - Disminución del tono muscular.
    - Alteraciones de la pauta respiratoria habitual: la risa no altera el ritmo respiratorio, sin embargo, sí induce cambios en el ciclo inspiración-espiración, determinando una mayor frecuencia espiratoria.
    - Espiración forzada.
    - Aumento del volumen de aire inspirado en cada ciclo.
    - Producción de movimientos sacádicos de baja amplitud y alta frecuencia, que se corresponden con el carcajeo.

- Expresión vocal:

Cuando una persona está alegre tiende a elevar el tono de la voz y a aumentar

su sonoridad, al tiempo que introduce un mayor número de variaciones tonales en su discurso. Cuando la alegría se manifiesta en forma de risa, se emiten una amplia variedad de sonidos, siendo los más característicos los que adoptan la forma sonora “ja-ja” o “je-je”.

Los efectos que produce la alegría sobre los diferentes parámetros vocales, en comparación con el estado neutro, son los siguientes:

- Fluencia:
  - El número de sílabas por segundo puede ser mayor o igual.
  - La duración de la sílaba puede ser menor o igual.
  - La duración de las vocales acentuadas puede ser mayor o igual.
  - Menor número y duración de las pausas.
- Frecuencia fundamental y prosodia:
  - Incremento del tono medio y su rango.
  - Mayor desviación de F0.
  - Mayor frecuencia de las sílabas acentuadas (incremento de la velocidad de locución).
  - Mayor gradiente de ascenso y descenso de F0.
- Esfuerzo vocal y tipo de fonación:
  - Incremento de la intensidad media.
  - Mayor desviación de la intensidad.
  - El gradiente de ascenso y descenso de la intensidad puede ser mayor o igual.

3.4.5.La emoción de tristeza

La tristeza es el sentimiento negativo caracterizado por un decaimiento en el

estado de ánimo habitual de la persona, que se acompaña de una reducción significativa en su nivel de actividad cognitiva y conductual, y cuya experiencia subjetiva oscila entre la congoja leve y la pena intensa propia del duelo o de la depresión.

Esta emoción se plantea ante situaciones que nos suponen alguna pérdida o que nos acarrean algún perjuicio o daño. Pero la tristeza no tiene por qué tener siempre un cariz negativo.

1.1.1.9.Desencadenantes de esta emoción

- Pérdida de una meta valiosa.
- Contingencia aversiva.
- Experiencia de otros.

1.1.1.10.Activación asociada a la tristeza

- Efectos subjetivos:
    - Vivencia desagradable.
    - Abatimiento, impotencia y aflicción.
    - Actitud pesimista.
    - Actitud reflexiva.
- Perfil psicofisiológico:
    - Actividad cardiovascular:
        - Ligero aumento de la frecuencia cardíaca.
        - Reducción del volumen de sangre bombeado al árbol arterial en cada latido.
        - Aumenta la resistencia vascular periférica.
        - Elevación de los niveles de presión sanguínea sistólica y diastólica.
    - Actividad electrodérmica:

- Aumenta el nivel de conductancia de la piel, alcanzando valores significativamente más altos que los observados con ocasión de otros estados emocionales.

- Actividad respiratoria y de la musculatura esquelética:
  - Elevación del tono muscular general. Cuando la intensidad de la tristeza aumenta (dando paso a estados próximos a la depresión), el efecto se invierte, induciendo una reducción en el nivel de tensión muscular.
  - El ritmo respiratorio se mantiene estable.
  - Aumento de la amplitud respiratoria.

- Expresión vocal:

Cuando nos sentimos tristes, se producen descensos en la media y el rango de la frecuencia fundamental de la señal vocal, y disminuye también su intensidad. Es decir, tanto la potencia como el ritmo o tasa media con la que vibran las cuerdas vocales, se atenúa sustancialmente bajo el influjo de este tipo de emoción. Nuestro tono de voz resulta más bajo y monótono, de menor sonoridad e intensidad. La fluencia verbal también se reduce, disminuyendo el número de palabras articuladas y ampliándose el tiempo necesario para su articulación. El habla se torna así cansina y lenta. Además, la tristeza puede asociarse también con cadencias tonales descendentes; esto es, el tono de voz tiende a disminuir progresivamente a lo largo de la pronunciación de la frase.

A continuación se muestran, de forma esquemática, los efectos que produce la tristeza sobre los diferentes parámetros vocales, en comparación con el estado neutro, son los siguientes:

- Fluencia:
  - Menor número de sílabas por segundo.
  - Mayor duración de la sílaba.
  - La duración de las vocales acentuadas puede ser mayor o igual.
  - Mayor número y duración de las pausas.

- Frecuencia fundamental y prosodia:
  - Tono medio más bajo que el normal.
  - Menor desviación de F0.
  - Estrecho rango de F0.
  - Menor frecuencia de sílabas acentuadas (velocidad de locución lenta).
  - Menor gradiente de ascenso y descenso de F0.
- Esfuerzo vocal y tipo de fonación:
  - La intensidad media puede ser menor o igual.
  - Menor desviación de la intensidad.
  - Menor gradiente de ascenso y descenso de la intensidad.

3.4.6.La emoción de enfado

El enfado es el sentimiento que surge ante una impresión desagradable, una molestia producida por alguna cosa (dicha o hecha). Se trata de un disgusto o enojo, generalmente contra otra persona.

1.1.1.11.Desencadenantes de esta emoción

- Situaciones frustrantes (obstrucción del acceso a una meta, transgresión de normas y derechos, extinción de contingencias aprendidas,…).
- Situaciones aversivas (inductores de dolor, situaciones desagradables de olor, frío, calor, ruido,…).

1.1.1.12.Activación asociada al enfado

Podemos distinguir dos tipos de enfado, el enfado el caliente y el enfado en frío, en los que la activación emocional tendrá características diferentes. El enfado en caliente se caracteriza por la sobre-articulación, así como por un elevado tono medio. El enfado en frío comparte muchas de las características del enfado en caliente, pero presenta un rango de F0 bastante menor. Las

características que veremos a continuación son las propias del enfado en caliente, siendo el enfado en frío un caso particular.

- Efectos subjetivos:
  - Vivencia desagradable.
  - Estado de alta activación.
  - Conducta poco reflexiva.
- Perfil psicofisiológico:
  - Actividad cardiovascular:
    - Aumenta la tasa cardíaca.
    - Aumenta la contractilidad miocardial.
    - Aumenta la presión sanguínea sistólica y diastólica.
    - Aumenta la resistencia vascular periférica.
  - Actividad electrodérmica:
    - Aumenta el nivel de conductancia de la piel y se incrementa el número de fluctuaciones espontáneas de la misma.
  - Actividad respiratoria y de la musculatura esquelética:
    - Aumenta el tono muscular, pudiendo ir acompañado de un nivel de tensión mayor en determinados grupos musculares.
    - Ritmo respiratorio más agitado y frecuente, mientras que la amplitud suele mantenerse en niveles basales.
  - Actividad endocrina:
    - Se segrega una mayor cantidad de adrenalina al torrente sanguíneo, lo que afecta, entre otras cosas, a la aceleración de la frecuencia cardíaca, al aumento de la fuerza muscular a la vasodilatación del árbol bronquial.

- Expresión vocal:

Los parámetros de intensidad y frecuencia de la expresión vocal cambian significativamente cuando nos sentimos malhumorados. Normalmente se produce un incremento tanto en la intensidad como en la frecuencia fundamental media de la señal de voz, es decir, aumenta el vigor y el ritmo o tasa media con la que vibran las cuerdas vocales.

La expresión vocal del enfado es la que muestra mayores diferencias entre los dos tipos de enfados que podemos tener. En el enfado en caliente aumentan bastante la frecuencia fundamental y su variabilidad, el habla se acelera (hay una mayor fluencia verbal) y se torna más enérgica. En cambio, en el enfado en frío, estas variaciones no se dan o aparecen mucho más atenuadas y localizadas en una parte concreta de la frase (por ejemplo, al final de ella).

De esta forma, los efectos del enfado en caliente sobre los diferentes parámetros vocales, en comparación con el estado neutro, son los siguientes:

- Fluencia:
  - El número de sílabas por segundo puede ser mayor o igual.
  - La duración de la sílaba puede ser mayor o igual.
  - Mayor duración de las vocales acentuadas.
  - Menor número y duración de las pausas.
- Frecuencia fundamental y prosodia:
  - Tono medio alto.
  - Mayor desviación de F0.
  - Amplio rango de F0.
  - Mayor frecuencia de sílabas acentuadas (velocidad de locución rápida).
  - Mayor gradiente de ascenso y descenso de F0.

- Esfuerzo vocal y tipo de fonación:
  - Mayor intensidad media.
  - Mayor desviación de la intensidad.
  - Mayor gradiente de ascenso y descenso de la intensidad.

3.4.7.La emoción de aburrimiento

El aburrimiento es un estado de desinterés o de falta de energía, como reacción a estímulos que se perciben como monótonos, repetitivos o tediosos. Se produce por la falta de cosas interesantes para ver, oír, etc., o para hacer cuando no se desea estar sin hacer nada.

Al tratarse de una emoción secundaria no se ha estudiado su activación emocional, como en el caso de las emociones anteriores, que eran primarias o básicas.

3.4.8.Comparativa de la expresión vocal de las distintas emociones

En la siguiente tabla se muestra una comparación entre los distintos valores que toman las emociones estudiadas (a excepción del aburrimiento), comparándolo con el estado neutro, para las características prosódicas: F0, rango de F0 y velocidad.

Tabla 1. Comparación de las características prosódicas de las emociones primarias.

	F0	Rango F0	Velocidad
Sorpresa	↑	↑	≈
Asco	↓	↑	↓
Miedo	↑	↑	↑
Alegría	↑	↑	↑
Tristeza	↓	↓	↓
Enfado (caliente)	↑	↑	↑

4. DESCRIPCIÓN DE LAS BASES DE DATOS

El objetivo de este capítulo es ofrecer una descripción detallada sobre las dos bases de datos (castellana y alemana) empleadas para evaluar el sistema desarrollado en este proyecto. Describiremos el tipo de ficheros de voz disponibles en cada una de ellas, así como el etiquetado empleado para los distintos ficheros. Finalmente, realizaremos una caracterización de la interpretación hecha por los locutores para cada una de las emociones.

4.1. SES (Spanish Emotional Speech database)

La base de datos SES fue grabada por un único actor profesional masculino. Las emociones interpretadas son: alegría, enfado, tristeza y sorpresa. También se dispone de grabaciones de voz interpretada según el estado neutro, para tomarlo como voz de referencia. El texto de los ficheros grabados no posee ningún contenido emocional intrínseco.

SES está compuesta por unos ficheros de audio en formato *pcm* y unos ficheros de marcas (*pmk*), que poseen información prosódica de las frases. Disponemos de dos tipos de fichero de audio, que explicamos a continuación:

- **Bloques:** cada bloque está formado por cuatro párrafos. Para cada una de las emociones se dispone de tres sesiones diferentes, excepto para la voz neutra de la que sólo disponemos de dos sesiones, por lo que el número total de ficheros que tenemos de este tipo es:

  Nº de bloques = (3 sesiones/emoción * 4 emociones) + (2 sesiones/voz_neutra * 1 voz_neutra) = **14 bloques**

  Nº de párrafos = 4 párrafos/bloque * 14 bloques = **56 párrafos**

- Frases del cuarto párrafo: en los párrafos el actor dispone de más tiempo y texto para transmitir la emoción interpretada, lo que hace que pueda haber partes muy similares a la voz neutra y puede suceder que a lo largo del párrafo haya partes que identifiquemos con una cierta emoción y otras con otra, eligiendo finalmente aquella que se dé con mayor probabilidad. Para conseguir unidades de menor duración que se identifiquen más fácilmente con una única emoción, vamos a dividir el cuarto párrafo en frases, de esta forma conseguiremos frases que tengan distintas componentes tanto segmentales como prosódicas, de las que puedan tener aquellas que se graben de forma independiente. Cada una de las sesiones que tenemos para cada emoción del párrafo cuarto contiene un total de 14 frases, por lo que al dividirlos obtenemos:

  Nº frases cuarto párrafo = 14 frases/párrafo * ((3 sesiones/emoción * 4 emociones) + (2 sesiones/voz_neutra * 1 voz_neutra)) = **196 frases**

- **Frases:** se trata de frases grabadas independientemente por el actor. Tenemos 15 frases diferentes y, al igual que en el caso de los párrafos, tenemos tres sesiones para cada una de las emociones y dos sesiones para la voz neutra. Por tanto, el número total de frases que tenemos es:

  Nº frases = 15 frases * ((3 sesiones/emoción * 4 emociones) + (2 sesiones/voz_neutra * 1 voz_neutra)) = **210 frases**

Por otra parte, disponemos de los ficheros de marcas, que tienen la siguiente información:

- Número de fonemas de la frase (denominados *segmentos* en el fichero de marcas).
- Número de periodos de la frase (denominados *unidades* en el fichero de marcas). Cada fonema estará formado por varios periodos.
- Instante de comienzo de cada periodo.
- Frecuencia de cada periodo.
- Transcripción fonética de cada fonema, indicando, en el caso de las vocales, si es

tónica o átona.

A continuación se muestra un ejemplo de parte de la información contenida en los ficheros de marcas:

```
NUM_SEGMENTS: 21
NUM_UNITS: 141
PIT: 0.213125,  78.05, n
PIT: 0.225938, 156.86, n
PIT: 0.232313, 134.45, n
PIT: 0.239750, 129.03, n
PIT: 0.247500, 152.38, n
PIT: 0.254063, 161.62, n
PIT: 0.260250, 175.82, n
PIT: 0.265938, 172.04, n
PIT: 0.271750, 161.62, 'o
PIT: 0.277938, 188.24, 'o
PIT: 0.283250, 202.53, 'o
             .
             .
             .
```

Figura 1. Información contenida en un fichero de marcas.

4.1.1.Etiquetado de los ficheros

Los ficheros de la base de datos SES los utilizaremos tanto para obtener los ficheros de características relacionadas con rasgos segmentales (MFCC - Mel Frequency Cepstral Coefficients), como para los relacionados con características prosódicas, pero sólo emplearemos las frases grabadas de manera independiente para obtener estas últimas características.

Los ficheros de SES de características relacionadas con rasgos segmentales están etiquetados con nombres cuya longitud es de 15 letras (en el caso de los ficheros cuya longitud es menor, rellenaremos con los ceros correspondientes para que el etiquetado de todos los ficheros tenga la misma estructura).

La extensión de cada uno de los ficheros será diferente en función de lo que contenga. En la siguiente tabla se muestra las posibles extensiones:

Tabla 2. Relación de extensiones de los diferentes tipos de ficheros de SES

Fuente	Extensión
Audio	pcm
Ficheros de marcas	pmk
MFCC	par
Prosodia	pro, phr, gfr

Como comentábamos anteriormente, disponemos de tres sesiones de cada una de las emociones y de dos sesiones de voz neutra. Las sesiones están numeradas de diferente manera en cada emoción, tal y como vemos en la siguiente tabla:

Tabla 3. Relación entre la etiqueta y el posible número de sesiones de cada emoción de SES.

Emoción	Etiqueta	N° Sesiones
Alegría	A	“03” “04” “05”
Enfado	E	“09” “10” “11”
Sorpresa	S	“12” “13” “14”
Tristeza	T	“06” “07” “08”
Neutro	N	“01” “02”

Distinguimos diferentes etiquetados dependiendo si se trata de los párrafos, las frases del cuarto párrafo o las frases grabadas de manera independiente. Vemos cada uno de estos etiquetados a continuación:

- **Párrafos:** la estructura del etiquetado de cada uno de los cuatro párrafos es la siguiente:

R_<emoción>_00<nº sesión>_0000P<nº párrafo>.<extensión>

Siendo:

- **<emoción>**: etiqueta de la emoción del párrafo según la Tabla 3.
- **<nº sesión>**: número de la sesión correspondiente al bloque del que está extraído el párrafo según la Tabla 3.
- **<nº párrafo>**: número del párrafo dentro del bloque. En la siguiente tabla vemos el texto de cada uno de los cuatro párrafos (apareciendo el cuarto

párrafo dividido en las 14 frases por las que está formado):

Tabla 4. Texto correspondiente a cada uno de los cuatro párrafos de SES.

N° Párrafo	Texto
1	Los participantes en el Congreso marcharon después a El Escorial. Se trasladaron allí en un amplio autobús, en el que un guía iba explicando los monumentos relevantes del recorrido. La visita al monasterio fue comentada por el mismo guía que debía saber mucho sobre El Greco, en cuyo cuadro "el martirio de San Mauricio" se extendió ampliamente; no debía ser igual su conocimiento del resto de los cuadros que componían la pinacoteca, sobre los cuales pasó como un rayo, dando lugar a sonrisas cómplices.
2	Sergio era un joven serio y trabajador que vivía cerca de la hospedería del Monasterio de Guadalupe, en las Villuercas, comarca perteneciente a la provincia de Cáceres. Se ganaba la vida vendiendo recuerdos alusivos a la Virgen Morenita, desde llaveros a platos con la imagen grabada en esmalte vidriado. Tenía un problema y era que su tiendecita era de mala construcción y estaba en una parte del pueblo muy empinada, fenómeno por otra parte normal en aquel lugar. Había mucho turismo en la zona. Sergio tuvo la mala suerte de perder su tienda en las últimas inundaciones, pues un corrimiento de tierras se la llevó por delante, con lo cual se le acabó su modo de vida.
3	Pablo estudiaba en la Universidad Politécnica de Madrid y estaba deseando regresar a Medellín; echaba de menos los productos de la matanza y los quesos frescos que hacía su abuela. Ya faltaba poco para las vacaciones; entonces volvería a las orillas del Guadiana, bajo los chopos. Su deseo era tan grande que a veces se le hacían años los pocos días que faltaban.

4	Frase 1	La vida diaria a menudo no es tan fácil, aunque estemos en el final del siglo veinte.
	Frase 2	Sobre todo cuando los dos en la pareja trabajan.
	Frase 3	Siempre hay que preguntarse si ya se cambió la ropa, si la puerta tiene el cerrojo o si tengo la llave en el bolsillo.
	Frase 4	Yo llevo al niño en el coche.
	Frase 5	Todos los días; al colegio.
	Frase 6	Pero ¿quién hace la compra?
	Frase 7	Al final de la semana todo se acaba.
	Frase 8	No queda fruta los viernes.
	Frase 9	Los sábados dejaron la cuenta al cero.
	Frase 10	Y los domingos, aunque te dices que vivirás una feliz experiencia, la cosa no es tan sencilla: el niño sale con sus amigos.
	Frase 11	¿Hay algún chico en la esquina?
	Frase 12	¿Se cayó en el jardín?
	Frase 13	Desde luego, siempre gozan de perfecta salud y yo estoy aquí preocupándome por nada.
	Frase 14	Definitivamente, vivir no es tan sencillo ni al final del siglo veinte.

- **<extensión>**: extensión correspondiente según la *Tabla 2.*

Un posible ejemplo de etiquetado de uno de los párrafos sería:

R_E_0009_0000P3.par ⇒ Fichero de características relacionado con rasgos segmentales, correspondiente al párrafo tercero extraído de un bloque cuya emoción interpretada es el enfado y la sesión es la “09”.

- **Frases del cuarto párrafo:** cada párrafo cuarto está formado por catorce frases, que etiquetaremos de la siguiente forma:

R_<emoción>_00<nº sesión>_P4_F<nº frase>.<extensión>

Siendo:

- **<emoción>**: etiqueta de la emoción del párrafo según la Tabla 3.

- **<nº sesión>**: número de la sesión correspondiente según la Tabla 3.
- **<nº frase>**: número de la frase dentro del párrafo según la *Tabla 4*.
- **<extensión>**: extensión correspondiente según la *Tabla 2.*

Un posible ejemplo de etiquetado de una de las frases del cuarto párrafo sería:

R_A_0004_P4_F13.par ⇒ Fichero de características relacionado con rasgos segmentales, correspondiente a la frase número 13 del cuarto párrafo extraído de un bloque cuya emoción interpretada es la alegría y la sesión es la "04".

- **Frases:** las frases grabadas independientemente tendrán el siguiente etiquetado:

F_<emoción>_<nº frase><nº sesión>.<extensión>

Siendo:

- **<emoción>**: etiqueta de la emoción del párrafo según la Tabla 3.
- **<nº frase>**: número de la frase según la siguiente tabla:

Tabla 5. Texto correspondiente a cada una de las frases independientes de SES.

N° Frase	Texto
1	No queda fruta los viernes.
2	¿Ya se cambió de ropa?
3	¿Hay algún chico en la esquina?
4	El final del siglo veinte.
5	¿La puerta tiene cerrojo?
6	Tengo la llave en el bolsillo.
7	¿Se cayó en el jardín?
8	¿Rompió la yema del huevo?
9	Gozan de perfecta salud.
10	Vivirás una feliz experiencia.
11	Dejaron la deuda al cero.
12	Le gusta mucho el gregoriano.
13	Yo llevo al niño en el coche.
14	Llegó la reina del puño cerrado.
15	Arrizabalaga dejará la reyerta.

- **<nº sesión>**: número de la sesión correspondiente según la Tabla 3.
- **<extensión>**: extensión correspondiente según la *Tabla 2.*

A continuación se muestran dos ejemplos: uno para un fichero de características relacionadas con rasgos segmentales y otro para un fichero con características relacionadas con rasgos prosódicos:

F_T_15080000000.par ⇒ Fichero de características relacionadas con rasgos segmentales, correspondiente a la frase independiente número 15 de tristeza de la sesión "08". Como comentábamos al principio de este apartado, los ficheros relacionados con rasgos segmentales de SES deben tener una longitud igual a 15, por lo que en este caso hemos tenido que rellenar con ceros el nombre del fichero.

F_N_0101.pro ⇒ Fichero de características relacionadas con rasgos prosódicos, correspondiente a la frase independiente número 1 de voz neutra de la sesión "01".

4.1.2.Caracterización de las emociones

En el *capítulo* vimos la descripción general de las emociones. En este apartado resumiremos las características particulares de cada una de las emociones de SES, estudiadas ampliamente en 1.1.1.45.

1.1.1.13.Análisis cualitativo

A continuación resumimos las características cualitativas de cada emoción de SES:

- **Enfado:** se trata de la emoción más destacada. En el *apartado 1.1.1.10* comentábamos que había dos tipos de enfados: el enfado en caliente y el enfado en frío. El actor interpreta una amenaza verbal que sugiere un enfado en frío. Se produce un gran esfuerzo vocal que se traduce en la presencia de una distorsión que lo hace muy característico.

- **Tristeza:** esta emoción se caracteriza por una cierta monotonía de tono, una velocidad de locución lenta, con pausas abundantes y largas.

- **Alegría:** desde un punto de vista prosódico, destaca la presencia de varios patrones entonativos en las frases de modalidad enunciativa, variando la posición del foco de una manera no sistemática, pudiéndose dar el caso de que, para un mismo texto pero en dos sesiones diferentes, la palabra realzada pase del principio al final de la frase. La voz alegre se caracteriza por la presencia de voz emitida sonriendo y por una claridad y brillantez notables.

- **Sorpresa:** se caracteriza por un tono medio muy elevado, alcanzando niveles de F0 superiores a los 250 Hz.

1.1.1.14.Análisis cuantitativo de las duraciones y el ritmo

Tras realizar un exhaustivo análisis sobre las duraciones y el ritmo de las distintas emociones en 1.1.1.45, llegamos a las siguientes conclusiones:

- La voz triste es la más lenta, siguiéndola muy de cerca, la voz enfadada.
- La voz alegre es la más rápida, presentando la menor duración de las pausas.

- El mayor alargamiento prepausa de las vocales finales es el de la tristeza.
- La influencia de la prosodia en la identificación del enfado simulado por el actor es casi nula.

1.1.1.15. Análisis cuantitativo de la entonación

Desde el punto de vista de la entonación, dividiremos la frase en tres zonas, mediante un modelo de picos y valles: hasta la primera tónica, entre la primera tónica y la última, y a partir de la última tónica. Analizamos estas zonas en cada una de las emociones.

- **Alegría:** su tono inicial es superior al neutro, debido a la primera tónica. El tono final es similar y no parece que transmita ninguna emoción. Las oraciones interrogativas presentan un tono final más elevado, aunque la última tónica es la que alcanza el máximo valor. La declinación es superior a la producida en el estado neutro, pero la diferencia es que el valor de la última tónica es mayor.

- **Tristeza:** presenta valores de F0 inferiores a la voz neutra, tanto en la primera tónica como en la última, teniendo un rango y una pendiente de declinación similares. En oraciones interrogativas el fonema final presenta menor F0. Los valles también presentan menor F0. Se trata de una voz homogénea y repetitiva.

- **Sorpresa:** presenta los mayores niveles de F0 observados, tanto en la primera como en la última tónica, presentando una pendiente de picos creciente. Las oraciones interrogativas se caracterizan por una fuerte subida entre la última tónica y el último fonema. Es importante la diferencia de valores que se da entre las últimas tónicas en oraciones enunciativas e interrogativas. Los valles son similares a los de la voz neutra, produciendo un gran contraste con los picos, que hace que tengan un gran rango.

- **Enfado:** presenta un tono muy plano con valores no muy alejados de la voz neutra. Las oraciones interrogativas se diferencian por no tener su valor máximo en el último fonema, sino en la última tónica.

La elección de las características prosódicas relacionas con el contorno de F0 y con el ritmo, utilizadas en los vectores de entrenamiento y clasificación de los experimentos de identificación que describiremos en el *capítulo 1.1.1.33*, viene justificada por los análisis cualitativos y cuantitativos realizados en este apartado, los cuales desvelan que esas van a ser las características que probablemente influyan más en la diferenciación de unas emociones frente a otras.

4.2.EMODB (Berlin Database of Emotional Speech)

La base de datos EMODB 1.1.1.45 fue grabada como parte de un proyecto de investigación, entre 1997 y 1999. Las grabaciones fueron llevadas a cabo en la Universidad Técnica de Berlín, en el departamento de Acústica Técnica.

EMODB se compone de las grabaciones de frases independientes interpretadas por diez actores. Las emociones transmitidas por los actores son: alegría, enfado, aburrimiento, tristeza, asco y miedo. También se dispone de grabaciones de voz interpretada según el estado neutro.

Las frases están grabadas en formato *wav* (16 kHz., 16 bit, mono), que transformaremos, antes de obtener los ficheros de características, en formato *pcm*, para tener las frases de ambas bases de datos en el mismo formato (ya que las frases de SES están en formato *pcm*). Además de los ficheros de audio, disponemos de los siguientes ficheros de etiquetas:

- *nombreFichero.lablaut:* contiene información sobre el instante inicial y la transcripción fonética de cada uno de los fonemas, en formato de audio (ESPS/ waves+).

- *nombreFichero.labsilb:* contiene información sobre el instante inicial y la transcripción fonética de cada una de las sílabas, en formato de audio (ESPS/ waves+).

- *nombreFichero.silb:* contiene información sobre el instante inicial y la transcripción fonética de cada una de las sílabas, en formato ASCII.

Los ficheros de EMODB sólo los utilizaremos para obtener ficheros de características relacionadas con rasgos segmentales, por lo tanto, las posibles extensiones que tenemos en EMODB son las que se muestran en la siguiente tabla:

Tabla 6. Relación de extensiones de los diferentes tipos de ficheros de EMODB

FUENTE	EXTENSIÓN
Audio	wav, pcm
Ficheros de etiquetas	lablaut, labsilb, silb
MFCC	par

El número total de frases del que dispone esta base de datos es de **535**, siendo el número de ficheros de cada emoción, así como el número de ficheros de cada locutor, no homogéneo, sino el que se muestra en la siguiente tabla (Tabla 7). El hecho de que el número de ficheros no sea homogéneo va a ser un problema a la hora de realizar los experimentos de identificación, ya que esto hace que el número de vectores de cada una de las emociones sea diferente, de forma que en el entrenamiento obtenemos algunos modelos basados en muchos datos y otros basados en menos, mientras que idealmente deberíamos obtener un modelo para cada una de las emociones basados en el mismo número de vectores de entrenamiento.

Tabla 7. Relación del número de frases de cada emoción y cada locutor en EMODB.

N° Locutor	Alegría	Enfado	Aburrimiento	Tristeza	Asco	Miedo	Neutro	TOTAL FICHEROS
3	7	14	5	7	1	4	11	49
8	11	12	10	9	0	6	10	58
9	4	13	4	4	8	1	9	43
10	4	10	8	3	1	8	4	38
11	8	11	8	7	2	10	9	55
12	2	12	5	4	2	6	4	35
13	10	12	10	5	8	7	9	61
14	8	16	8	10	8	12	7	69
15	6	13	9	4	5	8	11	56
16	11	14	14	9	11	7	5	71
TOTAL FICHEROS	71	127	81	62	46	69	79	**535**

4.2.1.Etiquetado de los ficheros

El etiquetado de las frases de la base de datos EMODB tiene la siguiente

estructura:

F_<emoción>_<n° locutor><código frase><nº sesión>.<extensión>

Siendo:

- **<emoción>**: etiqueta de la emoción de la frase, según la siguiente tabla:

Tabla 8. Etiqueta correspondiente a cada una de las emociones de EMODB.

Emoción	Etiqueta
Alegría (Freude)	F
Enfado (Wut)	W
Aburrimiento (Langeweile)	L
Tristeza (Trauer)	T
Asco (Ekel)	E
Miedo (Angst)	A
Neutro	N

- **<n° locutor>**: disponemos de grabaciones realizadas por diez locutores diferentes, numerados según se indica en la siguiente tabla (en la que indicamos si es hombre o mujer):

Tabla 9. Numeración y sexo de los locutores de EMODB.

N° Locutor	Sexo
03	Hombre
08	Mujer
09	Mujer
10	Hombre
11	Hombre
12	Hombre
13	Mujer
14	Mujer
15	Hombre
16	Mujer

- **<código frase>**: los códigos de las frases de las que disponemos en EMODB son los que se muestran en la siguiente tabla:

Tabla 10. Texto correspondiente a cada una de las frases de EMODB, junto con su código.

Código frase	Texto
a01	Der Lappen liegt auf dem Eisschrank
a02	Das will sie am Mittwoch abgeben
a04	Heute abend könnte ich es ihm sagen
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück
a07	In sieben Stunden wird es soweit sein
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen

- **<nº sesión>**: número de la sesión correspondiente. Puede ser: “a”, “b”, “c” o “d”.

- **<extensión>**: extensión correspondiente según la *Tabla 6.*

Un posible ejemplo de etiquetado de una de las frases de EMODB sería:

F_W_09b02c.par ⇒ Fichero de características relacionado con rasgos segmentales, correspondiente a la sesión “c” de la frase de enfado cuyo código es “b02” interpretada por el actor “09”.

4.2.2.Caracterización de las emociones

Las características específicas de las emociones grabadas por los actores en EMODB han sido estudiadas en 1.1.1.45, y se resumen a continuación:

- **Enfado:** se trata de la emoción en la que se produce mayor variación de F0. La distribución de valores va de +7 a –8 semitonos por segundo y se parece a una distribución gausiana. Una posible explicación de este amplio rango, es la estrategia del actor para expresar esta emoción: el actor pasa a un estado de enfado, y suelta toda su energía en una frase. En la base de datos hay tres tipos de frases en función de donde se sitúe la mayor energía: al principio, en medio o al final. Debido a los altos valores de la frecuencia fundamental,

pueden producirse tres tipos de fenómenos, en función del tipo de frase: una caída (si el pico está al principio de la frase), mantenerse (si está en medio) o una tendencia ascendente (si está al final).

- **Alegría:** la distribución de esta emoción va de –9 a +3 semitonos por segundo. Se producen muchas subidas importantes de la frecuencia fundamental.

- **Tristeza:** se trata de la emoción en la que se produce menos variación de F0. Tiene un estrecho rango de la frecuencia fundamental. Esto puede ser debido a que la tristeza está asociada a una mínima tensión de los músculos de la laringe. Como consecuencia del bajo esfuerzo vocal, el consumo de aire es pequeño, y la presión glotal (que es la principal causante de la declinación de la frecuencia fundamental) disminuye más despacio que en la voz neutra. Por lo general, los actores alcanzan la frecuencia fundamental más baja antes del final de la frase y siguen hablando con un tono chirriante. A veces aumenta F0 unos pocos hercios después de alcanzar el punto más bajo.

- **Miedo:** se producen –0,8 semitonos por segundo. Mientras que el enfado presenta la mayor variación de F0, y la tristeza y el asco muestran la mínima variación, la variación del miedo es moderada. Si suponemos que en la voz neutra, los factores fisiológicos causan la declinación de la frecuencia fundamental sobre la duración de la frase y que se producen –3 semitonos por segundo, podríamos pensar que no es posible evitar estas acciones al interpretar las emociones. Pero en las frases de miedo, se espera un mayor consumo de aire, que causa una frecuencia fundamental más alta. En las grabaciones de miedo, los actores a menudo hacen pausas cortas donde la inhalación es claramente audible. Si no hay pausas, los ruidos de inhalación son audibles después de la última sílaba.

- **Asco:** la distribución de valores de F0 es muy estrecha. El gradiente es más pequeño que en la voz neutra. Se caracteriza por grandes aumentos y disminuciones de la frecuencia fundamental.

- **Aburrimiento:** en esta emoción se produce la mayor tendencia descendente de F0 de todas las emociones. La distribución media es de –4 semitonos por segundo, alcanzándose hasta –10 semitonos por segundo. Podemos observar

que esta distribución es mucho mayor que la que se produce en la tristeza, aunque la naturaleza de ambas emociones sea similar. Los actores suspirar antes de comenzar la frase y el principio de éstas se caracteriza por un alto nivel de frecuencia fundamental. A lo largo de la frase, la frecuencia fundamental sufre una importante caída. La fuerte bajada de la frecuencia fundamental, junto con una expiración audible, hacen que el oyente distinga fácilmente la impresión de aburrimiento.

El análisis realizado sobre las emociones de EMODB nos permite seleccionar una serie de características prosódicas relevantes a la hora de reconocerlas, como pueden ser el valor medio de F0 o su rango. En un principio se abordó la identificación de las emociones de EMODB con características relacionadas con la prosodia, pero a la hora de dividir las frases en sus correspondientes grupos fónicos (de los cuales deberíamos extraer las características), nos dimos cuenta de que no teníamos la información lingüística necesaria para hacerlo, por lo que esto lo dejamos como una posible línea futura de investigación.

5. DEFINICIÓN DEL SISTEMA

En este capítulo vamos a realizar una descripción detallada de cada uno de los bloques en los que podemos dividir el sistema desarrollado para el reconocimiento automático de emociones a través de la voz, Estos bloque son: parametrización, normalización, entrenamiento y clasificación.

5.1. Diagrama de bloques

El diagrama de bloques que representa el sistema desarrollado es el siguiente:

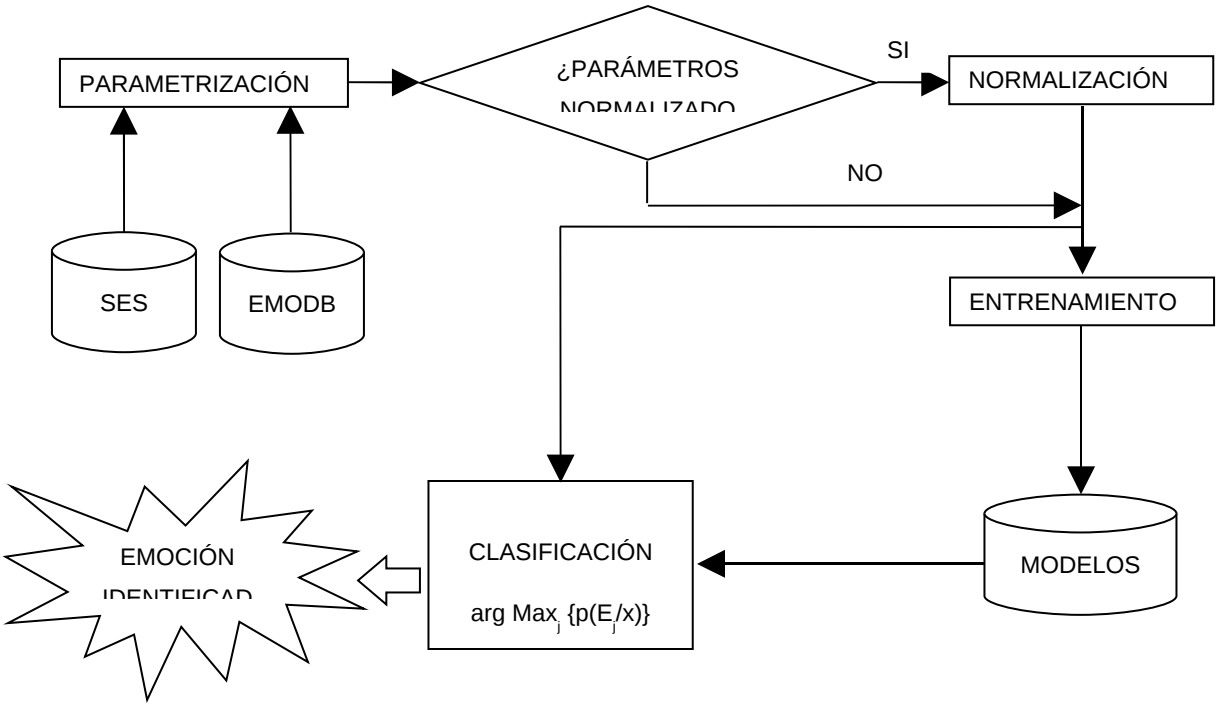


Figura 2. Diagrama de bloques del sistema implementado.

A continuación explicamos en detalle cada uno de los bloques representados en el diagrama anterior:

5.1.1.Parametrización

Este bloque nos va a permitir, a partir de los ficheros disponibles en cada una de las bases de datos (SES y EMODB), obtener ficheros formados por los vectores de características.

Llevaremos a cabo dos estrategias diferentes de parametrización, en función de con qué estén relacionados los vectores de características. Pueden estar relacionados con los rasgos segmentales (MFCC) o con la prosodia.

1.1.1.16.MFCC

Los MFCC (Coeficientes Cepstrales de las frecuencias de Mel – Mel Frequency Cepstral Coefficients) son coeficientes para la representación del habla basados en la percepción auditiva humana. Los MFCC muestran las características locales de la señal de voz asociadas al tracto vocal (dependiendo del instante de análisis), según el modelo filtro-fuente 1.1.1.45.

Los coeficientes cepstrales se derivan de la transformada de Fourier (FT - Fourier Transform) o de la transformada del coseno discreta (DCT - Discrete Cosine Transform), pero la particularidad básica es que en MFCC las bandas de frecuencia están situadas logarítmicamente, según la escala Mel, en la que el punto de referencia se define equiparando un tono de 1000 Hz., 40 dBs por encima del umbral de audición del oyente, con un tono de 1000 mels, tal y como se muestra en la siguiente figura:

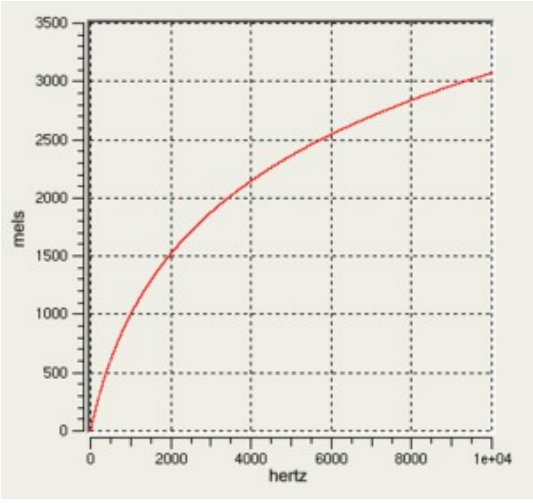


Figura 3. Gráfica mel-Hz.

Por encima de 500 Hz., los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente, de esta forma, cuatro octavas en la escala de hercios (por encima de 500 Hz.) se comprimen como dos octavas en la escala mel. De esta forma se modela la respuesta auditiva humana más apropiadamente que con las bandas espaciadas linealmente.

Los ficheros de los cuales vamos a extraer sus rasgos segmentales serán los disponibles tanto en SES como en EMODB. En SES disponemos de bloques formados por cuatro párrafos, que tendremos que dividir mediante el programa *Praat*. A su vez, el cuarto párrafo también lo segmentaremos en frases con *Praat*.

Los ficheros que contienen características segmentales (MFCC) están formados por el nombre del fichero, seguidos de la extensión *".par"*. Estos ficheros los obtenemos mediante el programa *Praat*, que nos permite obtener los coeficientes MFCC según los pasos indicados en el diagrama de bloques que aparece en la siguiente figura:

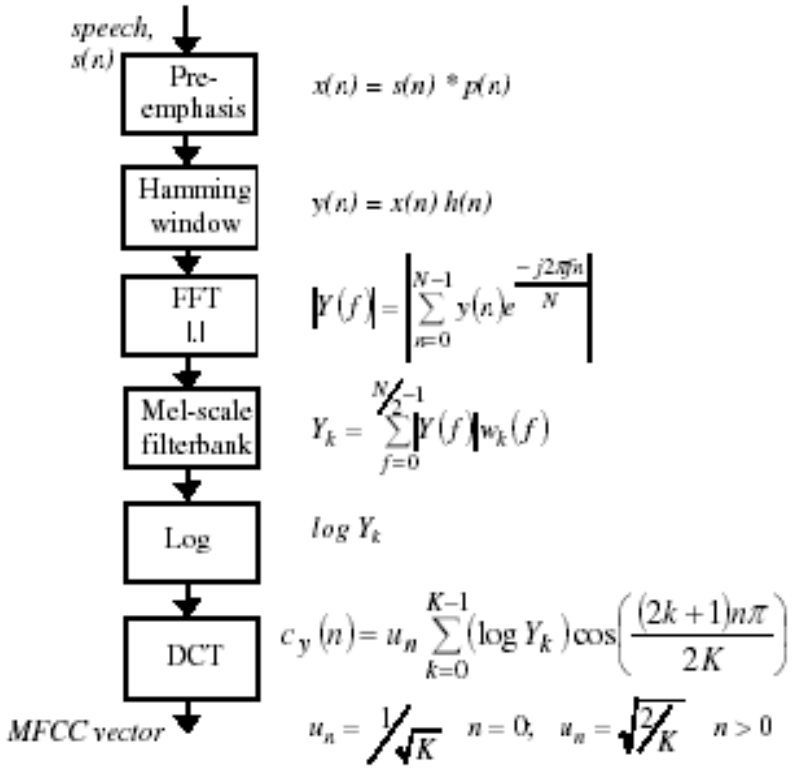


Figura 4. Método empleado por el programa *Praat* para calcular el vector de MFCC a partir de la señal de voz.

Tal y como aparece en la figura anterior, lo primero que debemos hacer es aplicar un preénfasis a la señal de voz, que aplana espectralmente dicha señal (realza las altas frecuencias), haciendo que su procesamiento sea menos susceptible a truncamientos. A continuación aplicamos una ventana de Hamming a la señal obtenida tras el preénfasis, para acotarla. El hecho de emplear una ventana de Hamming en vez de una rectangular, es que la rectangular da el máximo ajuste, pero produce grandes ondas laterales, mientras que la de Hamming no tiene tanta precisión frecuencial, pero provoca efectos mucho menores y es la utilizada de forma más común en el análisis del lenguaje. Una vez que tenemos nuestra señal enventanada, pasamos al dominio de la frecuencia mediante la FFT (Fast Fourier Transform). Calculamos los valores del banco de filtros distribuidos en frecuencia, según la escala Mel. Finalmente, calculamos el logaritmo de dicho banco, y mediante la DCT obtenemos los coeficientes cepstrales de las frecuencias de mel, que formarán el vector de MFCC.

Los parámetros que debemos especificar en la función del *Praat* que nos calcula los coeficientes MFCC son los siguientes:

- Número de coeficientes: en nuestro caso serán 12, debido a que éste es el número de coeficientes que se han utilizado ampliamente en reconocimiento automático de habla.
- Frecuencia de Nyquist: 8000 Hz.
- Duración de la ventana de análisis: 0,025 segundos.
- Desplazamiento: 0,01 segundos.

1.1.1.17.Prosodia

La segunda estrategia de clasificación que llevaremos a cabo será aquella en la que las características de cada fichero las obtengamos mediante rasgos prosódicos (entonación y ritmo).

La prosodia es una rama de la lingüística que analiza y representa formalmente aquellos elementos suprasegmentales de la expresión oral, que son elementos que afectan a más de un fonema y que no pueden segmentarse en unidades menores, tales como el acento, los tonos, el ritmo y la entonación. Su manifestación concreta en la producción de la palabra se asocia de este modo a las variaciones de la frecuencia

fundamental, de la duración y de la intensidad que constituyen los parámetros prosódicos físicos.

En la generación de modelos prosódicos se pueden abordar cuatro características básicas: la duración, la intensidad, las pausas y los movimientos melódicos. La entonación es básicamente la evolución de la frecuencia fundamental, mientras que el ritmo incluye tanto las duraciones de cada uno de los signos de síntesis como la localización y duración de las pausas. Las variables típicamente utilizadas en el análisis de la prosodia son, por ejemplo, el tipo de oración, la duración en tiempo, el número de sílabas del grupo entonativo, la distancia a la última sílaba acentuada, la categoría gramatical de la palabra, etc.

En la Figura 5 y la Figura 6 se muestra el fichero de audio (en la parte de arriba) junto con el contorno de F0 (en la parte de abajo) de un fichero de voz neutra y otro de sorpresa, respectivamente.

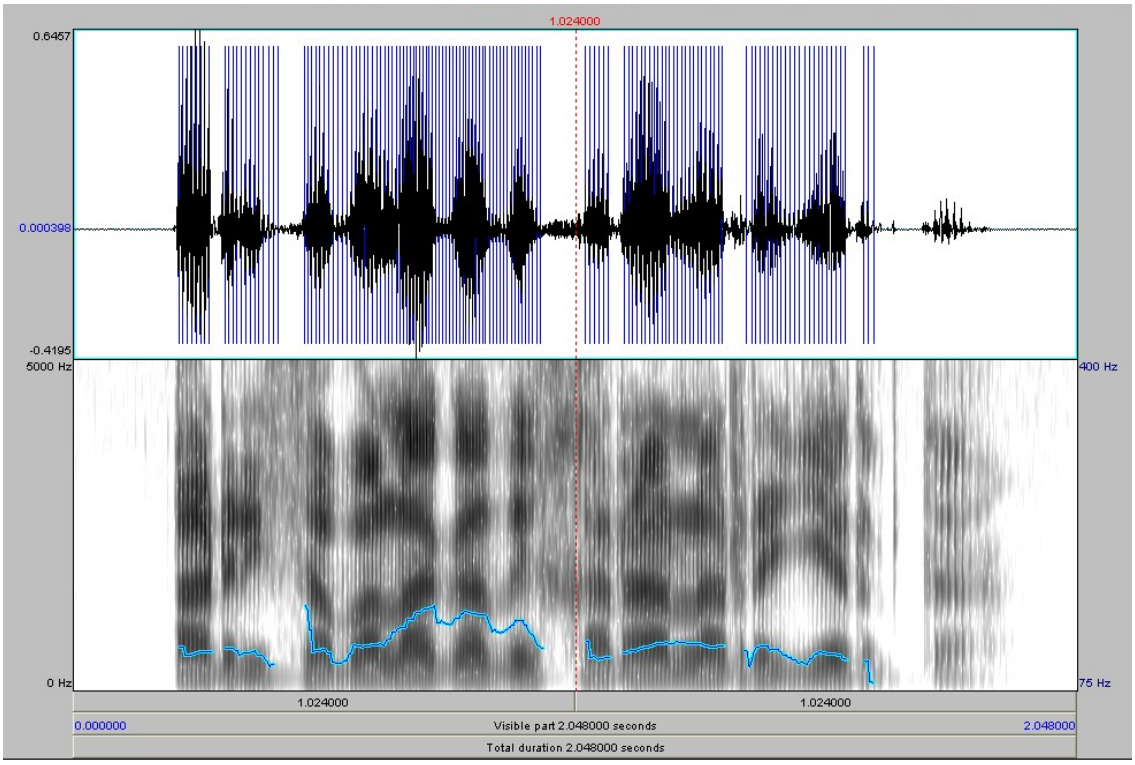


Figura 5. Representación con el *Praat* del fichero de audio *F_N_1502.PCM* junto con el contorno de F0.

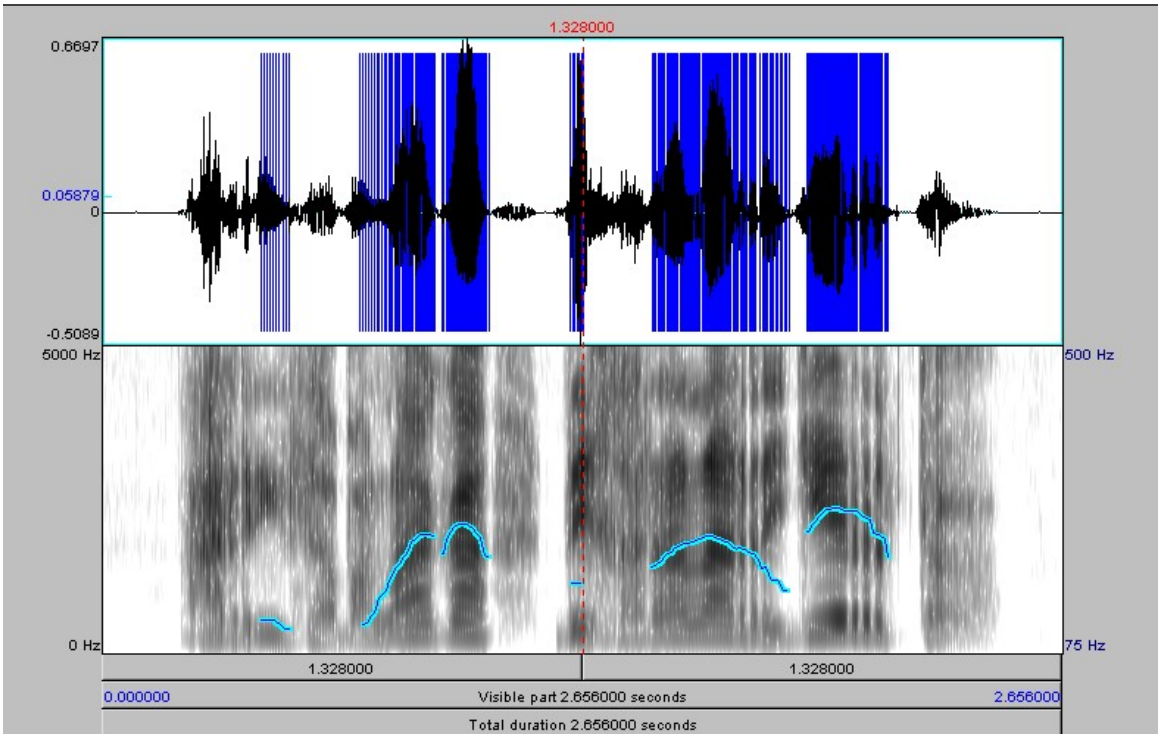


Figura 6. Representación con el *Praat* del fichero de audio *F_S_1512.PCM* junto con el contorno de F0.

En las figuras anteriores podemos observar como el contorno de F0 del fichero de sorpresa es menos homogéneo que el del fichero de neutra, teniendo la frecuencia fundamental un mayor rango y presentando valores mayores.

En general, la entonación española consta de una primera rama ascendente que comprende desde el primer sonido hasta el primer acento tónico. A partir de aquí se mantiene subiendo y bajando, hasta la parte del último acento. La elevación de esta última parte indica que la frase no esta completa. Su descenso indica la finalización de la frase, y la combinación de ambas, ascendente y descendente, que la frase es interrogativa.

Todos los experimentos realizados con características basadas en prosodia están hechos sobre las frases independientes de la base de datos SES. En SES disponemos del conocimiento lingüístico necesario para dividir las frases en grupos fónicos, a partir de los ficheros de marcas explicados en el apartado *1.1.1.12*. Dichos grupos los representaremos mediante un conjunto de características que serán las que utilicemos en el entrenamiento y la clasificación.

Los ficheros de características se extraen mediante el programa en linux *GeneraPRO.cpp*, a partir de los ficheros de marcas disponibles en SES. Dicho programa realiza los pasos que indicamos a continuación:

Lo primero de todo es crear una estructura para cada fonema de la frase, con los siguientes datos:

- Tiempo inicial: tiempo en el que comienza el fonema.
- Tiempo final: tiempo en el que finaliza el fonema.
- Duración: diferencia entre el tiempo final y el inicial.
- Mínimo: valor mínimo de la frecuencia fundamental dentro del fonema.
- Máximo: valor máximo de la frecuencia fundamental dentro del fonema.
- Valor medio de la frecuencia fundamental: muestra el valor medio que toma la frecuencia fundamental a lo largo del fonema.
- Tag: trascripción del fonema.

Una vez que hemos caracterizado cada uno de los fonemas, debemos buscar las tónicas de la frase, que representarán los máximos del contorno de F0.

También debemos localizar los valles. Un valle es un mínimo entre dos picos del contorno de F0. Los valles no pueden estar constituidos por consonantes sordas (que son: “p”, “f”, “t”. “z”, “s”, “c”, “k” y “x”), ya que el valor de frecuencia es muy bajo, porque se trata de un valor artificial, al que se le asigna un mínimo.

Debemos tener en cuenta que el valle tampoco puede estar formado por la segunda vocal de un diptongo o hiato, es decir, si encontramos una tónica y el siguiente fonema es una vocal, este no puede constituir un valle, tendremos que seguir buscando hasta encontrar el siguiente mínimo de F0.

Cuando hemos localizado todos los valles y los picos de la frase, lo que hacemos es segmentar las frases en grupos fónicos. Un grupo fónico va desde un valle hasta el siguiente, teniendo en cuenta que el primer grupo fónico va desde el inicio de la frase hasta el primer valle y el último grupo fónico va desde el último valle hasta el final de la frase.

En la siguiente figura (Figura 7) se muestra la división de un fichero en sus respectivos grupos fónicos. El fichero elegido es el *F_A_0103.PMK*, cuya frase interpretada es:*"No queda fruta los viernes".* En la parte de arriba de la figura aparece la representación del fichero de audio, a continuación aparece el contorno de F0 sin interpolar e interpolado. Y por último, tenemos la segmentación del contorno de F0 interpolado en los distintos grupos fónicos en los que podemos dividir dicha frase.

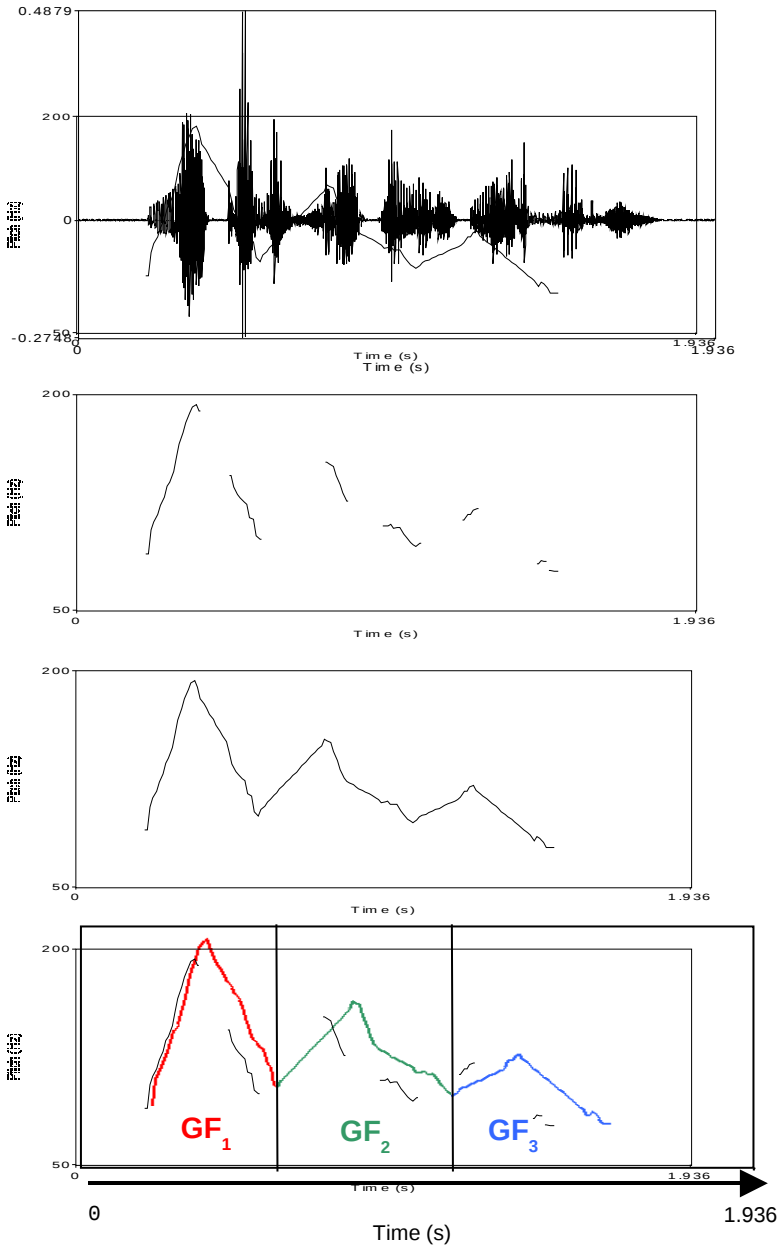


Figura 7. Representación del fichero de audio *F_A_0103.PMK*, su contorno de F0 sin interpolar e interpolado y segmentación en grupos fónicos.

Para cada grupo fónico creamos una estructura con los siguientes datos:

- Inicio: representa la posición inicial del grupo fónico.
- Fin: representa la posición final del grupo fónico.
- Mínimo: representa el valor mínimo de la frecuencia fundamental dentro del grupo fónico.
- Máximo: representa el valor máximo de la frecuencia fundamental dentro del grupo fónico.
- Rango: muestra la máxima variación de la frecuencia fundamental a lo largo del grupo fónico. Es decir, es la diferencia entre el máximo y el mínimo de frecuencia (los dos valores anteriores).
- Valor medio de la frecuencia fundamental: muestra el valor medio que toma la frecuencia fundamental a lo largo de todo el grupo fónico en cuestión.
- Pendiente de subida: es la pendiente desde el valle inicial hasta el pico del grupo fónico en cuestión.
- Pendiente de bajada: es la pendiente desde el pico hasta el valle final del grupo fónico en cuestión.

Adicionalmente calcularemos dos características relacionadas con el ritmo, que posteriormente emplearemos en nuestro sistema para el reconocimiento de emociones. Estas características son:

- Velocidad de locución de la frase: se define como el número de fonemas de la locución dividido entre su duración.

$$v_{locución} = \frac{n^{o}\ fonemas}{duración_{locución}}$$

- Velocidad de cada grupo fónico: la calcularemos de la misma forma que calculamos la velocidad de locución de la frase, pero particularizando para un grupo fónico, es decir, es el número de fonemas del grupo fónico dividido entre la duración dicho grupo.

$$v_{GF} = \frac{n^{o}\ fonemas_{GF}}{duración_{GF}}$$

Para los experimentos basados en prosodia, vamos a tener tres tipos diferentes de ficheros de características por cada fichero disponible en la base de datos. Todos ellos estarán formados por el nombre del fichero de voz que vamos a parametrizar seguido de una de las extensiones que explicamos a continuación. La primera (*".pro")* está relacionada con la frecuencia fundamental, y las dos últimas (*".phr"* y *".gfr")* están relacionadas con el ritmo.

- *".pro":* ficheros dónde los vectores de características contendrán los siguientes datos de cada grupo fónico: valor medio de F0, valor mínimo de F0, valor máximo de F0, rango de F0, pendiente de subida y pendiente de bajada. También generaremos otro grupo de ficheros cuyo vector de características contenga los mismos datos que el anterior, salvo el valor medio de F0. Estos ficheros tendrán la misma extensión pero se guardarán en una carpeta diferente.
- *".phr":* ficheros que contienen únicamente la velocidad de locución de la frase.
- *".gfr":* ficheros cuyos vectores de características contiene un único dato que es el valor de la velocidad de cada grupo fónico.

Realizaremos distintos experimentos en función de los grupos fónicos analizados: tendremos un experimento en el que emplearemos todos los grupos fónicos, y otros en los que sólo usemos los grupos fónicos iniciales, los finales o los medios. Por tanto, dentro de los ficheros cuya extensión sea *".pro"* y *".gfr"* distinguiremos cuatro tipos de ficheros:

- *"nombreFichero.pro"* o *"nombreFichero.gfr"*: contienen las características de todos los grupos fónicos.
- *"nombreFichero_Ini.pro"* o *"nombreFichero_Ini.gfr"*: contienen sólo las características del grupo fónico inicial de cada frase.
- *"nombreFichero_Med.pro"* o *"nombreFichero_Med.gfr"*: contienen sólo las características de los grupos fónicos medios de cada frase.

- *“nombreFichero_Fin.pro”* o *“nombreFichero_Fin.gfr”*: contienen sólo las características de los grupos fónicos finales de cada frase.

5.1.2.Normalización

En este proyecto sólo hemos abordado la normalización de los parámetros segmentales, dejando como propuesta el estudio de la parametrización de características prosódicas.

En los ficheros de EMODB, dada la existencia de 10 locutores diferentes, tendremos una variabilidad asociada a las diferencias acústicas relacionadas con las características propias de la voz de cada locutor.

El conjunto de diferencias acústicas o distorsiones no correspondientes a las propias diferencias entre locutores, se denomina “variabilidad del canal”, y en él se encuadran distintas distorsiones producidas por diferentes canales, tipos de micrófono o condiciones ambientales, entre otros. Se ha demostrado que el uso de técnicas de compensación de canal, ya sea sobre el audio, los parámetros a modelar o el propio modelo, mejora las tasas de reconocimiento.

En nuestro caso, aplicaremos técnicas de normalización sobre las características obtenidas a partir de los ficheros de voz.

Podemos realizar distintos tipos de normalización:

- CMN (Cepstral Mean Normalization): es una normalización basada en coeficientes cepstrales. Consiste en dividir una locución en cortas ventanas de tiempo y extraer de ellas un cierto número de coeficientes cepstrales. Para cada uno de ellos, extraeremos la media del coeficiente a lo largo de toda la locución. Restaremos la media obtenida a cada uno de los coeficientes cepstrales, como se indica en la siguiente fórmula:

$$\bar{c}_i^t = c_i^t - \mu_{c_i}$$

Siendo:

$\mu_{c_i} = \frac{\sum_{t=1}^{n} c_i^t}{n}$ ≡ media de los coeficientes cepstrales.

c_i^t ≡ valor del coeficiente cepstral *i* dada la trama *t*.

$\overline{c}_i^t$ ≡ valor normalizado respecto a la media del coeficiente cepstral.

n ≡ número total de tramas.

De esta forma pretendemos reducir la distorsión introducida por elementos de variación lenta (como, por ejemplo, el ruido estacionario). Es decir, con esta normalización podemos compensar el canal.

- CVN (Cepstral Varianze Normalization): se trata de una normalización similar a la anterior en la que en lugar de calcular la media de los coeficientes cepstrales, calculamos la varianza. Luego dividimos cada uno de los coeficientes entre la raíz cuadrada de dicha varianza, según la siguiente fórmula:

$$\overline{c}_i^t = \frac{c_i^t}{\sqrt{\sigma_{c_i}^2}}$$

Siendo:

$\sigma_{c_i}^2 = \frac{\sum_{t=1}^{n} (c_i^t - \mu_{c_i})^2}{n}$ ≡ varianza de los coeficientes cepstrales.

$\overline{c}_i^t$ ≡ valor normalizado respecto a la varianza del coeficiente cepstral.

Gracias a esta normalización podemos compensar las características propias de cada locutor.

- CMN/CVN: consiste en combinar las dos técnicas anteriores.

$$\bar{c}_i^t = \frac{c_i^t - \mu_{c_i}}{\sqrt{\sigma_{c_i}^2}}$$

Siendo:

$\bar{c}_i^t$ ≡ valor normalizado respecto a la media y la varianza del coeficiente cepstral.

- Respecto a la voz del locutor: normalizamos el número total de ficheros de un determinado locutor respecto a parámetros obtenidos a partir de todos los ficheros disponibles para dicho locutor. Lo que buscamos con esta normalización es eliminar información no discriminativa del locutor que varía de forma no controlada entre las distintas locuciones.

- Respecto a la voz neutra: obtendremos los parámetros de normalización a partir de los ficheros de la voz neutra de cada locutor. Con este tipo de normalización lo que pretendemos es quedarnos con las características propias de cada emoción, independientemente de cómo sea la voz neutra de cada locutor.

1.1.1.18. Etiquetado de los ficheros normalizados

Debemos cambiar el nombre de los ficheros que normalicemos, para saber que las características que contiene dicho fichero están normalizadas. Para que todos los nombres de los ficheros normalizados tengan la misma estructura, añadiremos al nombre del fichero de características que teníamos sin normalizar las siguientes etiquetas, en función de la normalización que usemos, justo delante de la extensión del fichero:

- *"_CMN"*: para la normalización CMN.
- *"_CVN"*: para la normalización CVN.
- *"_MyV"*: para la normalización CMN/CVN.

Guardaremos en carpetas diferentes las normalizaciones respecto al locutor y respecto a la neutra.

5.1.3.Entrenamiento del sistema

Una vez generados los ficheros que caracterizan cada fichero de voz, lo que tenemos que hacer es entrenar las distintas emociones, para obtener un modelo para cada una de ellas, que será una mezcla de gausiana (GMM) 1.1.1.45.

El algoritmo GMM es un modelo paramétrico utilizado clásicamente en muchas técnicas de reconocimiento de habla como un modelo genérico probabilístico de varias densidades capaz de representar densidades arbitrarias. Se trata de modelos estadísticos en los que la estructura de secuencialidad temporal subyacente en la señal de habla queda eliminada.

Este modelo asume que la distribución de probabilidad de los parámetros observados toma la siguiente forma paramétrica:

$$p(x) = \sum_{i=1}^{m} \alpha_i N(x; \mu_i, \Sigma_i)$$

Donde:

- N (x; μ, Σ) representa la distribución normal p-dimensional, siendo μ el vector media y Σ la matriz covarianza definida como:

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right]$$

- Los términos α_i representan los pesos escalares positivos normalizados, cumpliéndose:

$$\sum_{i=1}^{m} \alpha_i = 1 \text{ y } \alpha_i \geq 0$$

Un supuesto fundamental del algoritmo GMM es declarar que los vectores de observación $\{x_t\}$ son independientes unos de otros. Esta simplificación del modelo, lo hace válido cuando el aspecto secuencial de la observación (en nuestro caso el índice de tiempo *t*) es irrelevante. De esta forma, podemos considerar el algoritmo GMM como una simplificación de los modelos ocultos de Markov (HMM – Hidden Markov Models) con distribuciones gausianas en las que todos los estados se unen y todas las probabilidades de transición que conducen a un estado dado, son iguales. En nuestro caso, la opción de utilizar GMM está justificada porque lo que nos interesa son las funciones de conversión segmentales, para las cuales la envolvente del índice de tiempo *t* sólo depende ese mismo instante de tiempo.

Una de las principales razones por las que usamos un modelo GMM es por su capacidad para proporcionar un suavizado entre varias distribuciones gausianas unimodales ($N(x; \mu, \Sigma)$). Cuando usamos el espectro de habla, las componentes del modelo GMM son clases acústicas que, muchas veces, representan varios acontecimientos fonéticos. Cada clase acústica está descrita por su centro, que es el vector media μ_i, así como por la característica que se extiende alrededor del centro, que viene dada por la matriz de covarianza Σ_i. La mezcla de pesos $\{\alpha_i\}$ representa la frecuencia estadística de cada clase en las observaciones.

5.1.4.Clasificación

Vamos a llevar a cabo dos estrategias de clasificación:

- Clasificación atendiendo a las características relacionadas con rasgos segmentales (MFCC).

- Clasificación atendiendo a las características relacionadas con la prosodia.

Los ficheros empleados en la clasificación serán diferentes a los empleados en el entrenamiento. Para decidir con que emoción se corresponde cada uno de los ficheros que tengamos que clasificar, debemos comparar las características de estos ficheros con los modelos de las distintas emociones obtenidos en el entrenamiento.

A partir de dichos modelos, mediante un clasificador bayesiano, estimamos la probabilidad de que un segmento de voz pertenezca a cada una de las emociones que

forman parte del sistema.

Dada una observación x_t de un segmento de voz parametrizado ($x = \{ x_0, x_1,..., x_t,..., x_{T-1}\}$), definimos la probabilidad de que dicha observación corresponda a una emoción E_j perteneciente al conjunto $E = \{ E_0, E_1,..., E_j,..., E_{J-1}\}$ como:

$$p(E_j / x_t) = \frac{p(x_t / E_j) \cdot P(E_j)}{P(x_t)} = \frac{p(x_t / E_j) \cdot P(E_j)}{\sum_k p(x_t / E_k) \cdot P(E_k)}$$

Suponiendo que las probabilidades de las emociones son equiprobables, es decir, $P(E_j) = P(E_k)$, nos queda:

$$p(E_j / x_t) = \frac{p(x_t / E_j) \cdot P(E_j)}{\sum_k p(x_t / E_k) \cdot P(E_k)} = \frac{p(x_t / E_j)}{\sum_k p(x_t / E_k)}$$

Por tanto, la probabilidad de que el segmento de voz *x* se corresponda con la emoción E_j será:

$$p(E_j / x) = \prod_{t=0}^{T-1} p(E_j / x_t) = \prod_{t=0}^{T-1} \frac{p(x_t / E_j)}{\sum_k p(x_t / E_k)}$$

Una vez que hemos calculado las probabilidades de que el segmento de voz se corresponda con cada una de las emociones, la emoción reconocida será:

$$E_{rec} = \arg Max_j \left\{ p\left(E_0 / x\right), p\left(E_1 / x\right),..., p\left(E_j / x\right),..., p\left(E_{J-1} / x\right) \right\}$$

A la hora de tomar la decisión, el denominador de cada una de las probabilidades de cada emoción es idéntico, es decir, no nos aporta información para tomar la decisión, por lo que podemos prescindir de él en el cálculo de las probabilidades para cada emoción.

6. IDENTIFICACIÓN DE EMOCIONES BASADA EN INFORMACIÓN SEGMENTAL

En este capítulo se definen y analizan los resultados obtenidos en los experimentos de identificación de emociones basados en información segmental, sobre las bases de datos SES y EMODB. La información segmental de los datos la extraeremos mediante los MFCC, que fueron explicados con detalle en el *apartado 1.1.1.16.*

Dentro de los experimentos realizados con los ficheros con características relacionadas con rasgos segmentales, podemos distinguir tres tipos: los realizados con la base de datos SES, los realizados con la base de datos EMODB y los experimentos con distintos idiomas, que son aquellos en los que entrenamos con datos de SES y clasificamos datos de EMODB y viceversa. En cada uno de ellos, se extraerán resultados según el número de gausianas utilizadas en el modelo. Para estos experimentos utilizaremos de una a cinco gausianas. El criterio que nos ha llevado a elegir cinco gausianas como máximo es el escaso volumen de datos disponibles para el entrenamiento de nuestro sistema.

6.1. Identificación con SES

Dentro de los experimentos realizados con los datos de SES, podemos distinguir dos tipos: aquellos en los que no hemos empleado ningún tipo de normalización de los vectores de características y aquellos en los que hemos normalizado dichas características según alguna de las normalizaciones explicadas anteriormente (*capítulo 1.1.1.17*).

6.1.1. Experimentos de identificación de emociones sobre SES sin normalización de características

En este apartado vamos a describir y analizar los resultados de los experimentos de identificación de emociones realizados con características, sin normalizar, relacionadas con los rasgos segmentales extraídos a partir de los datos de SES.

1.1.1.19. Descripción de los experimentos

Los cinco tipos de experimentos de identificación realizados se diferencian en el tipo de ficheros empleados en el entrenamiento y la clasificación, que pueden ser: frases, párrafos o una combinación de ambos tipos. Al entrenar con párrafos tenemos un mayor número de vectores de entrenamiento, lo que nos permite obtener unos modelos para cada una de las emociones más robustos. Algunos de los experimentos descritos a continuación serán independientes de texto (es decir, el texto de los datos de entrenamiento es diferente al de los datos de clasificación) y otros serán dependientes de texto.

En la siguiente tabla podemos observar el número de vectores de entrenamiento y clasificación disponibles para los experimentos que describimos en este apartado:

Tabla 11. Número de vectores de entrenamiento y clasificación para cada uno de los experimentos realizados sobre los datos sin normalizar de SES.

	N° de vectores de entrenamiento	N° de vectores de clasificación
EXPERIMENTO 1	234.175 (84%)	43.338 (16%)
EXPERIMENTO 2	151.453 (56%)	118.865 (44%)
EXPERIMENTO 3	182.066 (67%)	90.834 (33%)
EXPERIMENTO 4	43.338 (16%)	234.175 (84%)
EXPERIMENTO 5	118.865 (44%)	151.453 (56%)

A continuación describimos los cinco experimentos realizados sobre los datos sin normalizar de la base de datos SES, en función del tipo de ficheros empleado en el entrenamiento y la clasificación:

- EXPERIMENTO 1: *Entrenamiento con los párrafos y clasificación con las frases independientes.*

Entrenaremos nuestro sistema con los 56 párrafos y luego clasificaremos las 210 frases disponibles.

Como podemos observar en la Tabla 11, el número de vectores de entrenamiento disponible es 234.175 (84% de los datos), mientras que el número de los vectores que vamos a clasificar es de 43.338 (16% de los datos). Observamos que el número de vectores de entrenamiento es un orden de magnitud mayor que el de vectores de clasificación, siendo este caso en el que disponemos de un mayor número de datos para obtener los modelos.

Debemos tener en cuenta que este experimento es dependiente de texto, es decir, que entrenamos al sistema con frases que luego utilizamos en la clasificación, ya que, como hemos comentado en la descripción de la base de datos SES (*capítulo 1.1.1.12*), el texto de las frases independientes está contenido en el texto del párrafo cuarto.

- EXPERIMENTO 2: *Entrenamiento con los tres primeros párrafos y clasificación con las frases independientes y las frases del cuarto párrafo.*

Como las frases están sacadas del cuarto párrafo, entrenaremos con todos los párrafos de cada emoción menos ese y luego clasificaremos las frases del cuarto párrafo (de cada emoción también) y las frases independientes que tenemos. De esta forma conseguimos un experimento independiente de texto, es decir, en el entrenamiento no se han visto las frases que luego utilizaremos en la clasificación.

Observando la Tabla 11 vemos que el número de vectores de entrenamiento y clasificación en este caso es similar: tenemos 151.453 (56% de los datos) vectores de entrenamiento frente a 118.865 (44% de los datos) vectores de clasificación.

- EXPERIMENTO 3: *Entrenamiento con dos sesiones de las frases independientes y los tres primeros párrafos y clasificación con la otra sesión de las frases independientes y las frases del cuarto párrafo.*

Disponemos de tres sesiones diferentes para cada una de las emociones interpretadas por el actor al grabar las frases independientes. Utilizaremos dos de ellas para el entrenamiento y la otra para la clasificación. Además de las frases, también entrenamos con los tres primeros párrafos de cada bloque. Para la clasificación emplearemos la sesión que nos queda de las frases independientes y las frases del cuarto párrafo. Es decir, en este experimento utilizamos frases y párrafos en el entrenamiento. Además, al utilizar las frases independientes en el entrenamiento, hacemos que este experimento no sea independiente de texto.

Podemos dividir este experimento a su vez en tres, dependiendo de las sesiones que utilicemos para entrenamiento y para clasificación. Los resultados obtenidos son similares en los tres experimentos, por lo que obtendremos la media de ellos, para comparar los resultados con el resto de experimentos.

Según observamos en la Tabla 11, el número de vectores de entrenamiento es de unos 182.000 y el número de vectores de clasificación de unos 91.000, es decir, empleamos una proporción del 67% del total de ficheros para entrenamiento y el 33% restante para clasificación, que es una proporción adecuada.

- EXPERIMENTO 4: *Entrenamiento con las frases independientes y clasificación con los párrafos.*

Entrenaremos nuestro sistema con las 210 frases independientes y luego clasificaremos los 56 párrafos disponibles.

Este experimento es el inverso al *experimento 1*, por lo que en este caso tenemos muchos más vectores de clasificación que de entrenamiento (ver Tabla 11). De esta forma, los modelos que obtenemos para clasificar se basan en pocos datos, por lo que en principio no serán muy fiables.

- EXPERIMENTO 5: *Entrenamiento con las frases independientes y las frases del cuarto párrafo y clasificación con los tres primeros párrafos.*

Entrenamos nuestro sistema con las 210 frases independientes que tenemos y con las frases extraídas de los cuartos párrafos y luego clasificamos los tres primeros párrafos. Este experimento, al igual que el *experimento 2*, que es su inverso, también es independiente de texto.

En este caso, la unidad de clasificación es el párrafo, no la frase. Esto va a hacer que posiblemente la tasa de identificación para este experimento sea mejor debido a que cada uno de los ficheros que tenemos que clasificar estarán formados por muchos vectores de características, de forma que es más fácil que a lo largo del párrafo se identifiquen los distintos vectores con la emoción correcta, y en global, el párrafo se identifique. A pesar de esto, como observamos en la Tabla 11, el número de vectores de entrenamiento es ligeramente inferior al número de vectores de clasificación y debería ser al revés.

En el experimento anterior también teníamos el párrafo como unidad de clasificación, pero dado que teníamos muy pocos vectores de entrenamiento, los modelos obtenidos para cada emoción se presume que estarán poco entrenados y, por tanto, la tasa de identificación obtenida será posiblemente menor.

En la siguiente tabla se muestra un resumen con las diferencias entre los cinco experimentos descritos anteriormente:

Tabla 12: Diferencias entre los experimentos realizados con la base de datos de SES con los vectores sin normalizar.

	DEPENDIENTE DE TEXTO	ENTRENAMIENTO	CLASIFICACIÓN
EXP 1	Sí	Párrafos (1, 2, 3 y 4)	Frases
EXP 2	No	Párrafos (1, 2 y 3)	Frases + Frases del párrafo 4
EXP 3	Sí	Frases (2 sesiones) + Párrafos (1, 2 y 3)	Frases (1 sesión) + Frases del párrafo 4
EXP 4	Sí	Frases	Párrafos (1, 2, 3 y 4)
EXP 5	No	Frases + Frases del párrafo 4	Párrafos (1, 2 y 3)

1.1.1.20. Resultados de los experimentos

En la Tabla 13 se muestra el número total de ficheros identificados, así como la tasa de identificación y la banda de fiabilidad para los cinco experimentos descritos en el apartado anterior, realizados con los vectores sin normalizar, obtenidos a partir de los datos de SES. Se muestran los resultados en función del número de gausianas utilizadas en el modelo. En la última fila se muestra la mejor tasa de identificación, dado un cierto número de gausianas, apareciendo sombreada la casilla correspondiente al experimento con el que se consigue dicha tasa de identificación. En la última columna aparece el número total de ficheros de clasificación disponible en cada uno de los experimentos.

La banda de fiabilidad nos va a permitir saber si los resultados obtenidos para un experimento son mejores que otros de forma estadísticamente fiable. Es decir, si estamos comparando la tasa de identificación de dos de los experimentos, para un número concreto de gausianas, sabremos que esos resultados son estadísticamente fiables si se cumple lo siguiente:

Tasa_de_identificación_1 + Banda_de_fiabilidad_1 < Tasa_de_identificación_2 – Banda_de_fiabilidad_2

Siendo: Tasa_de_identificación_1 < Tasa_de_identificación_2

Vemos un ejemplo numérico para que quede más claro:

Si tenemos los siguientes datos:

- Tasa_de_identificación_1 = 87%
- Banda_de_fiabilidad_1 = 2,3%
- Tasa_de_identificación_2 = 95%
- Banda_de_fiabilidad_2 = 3,8%

Para que sea estadísticamente fiable se tiene que cumplir que:

$$87\% + 2{,}3\% < 95\% - 3{,}8\%$$

$$89{,}3\ \% < 91{,}2\%$$

Observamos que se cumple la desigualdad, por lo que los resultados son estadísticamente fiables.

Al comparar los resultados obtenidos en los diferentes experimentos, deberemos comprobar que se trata de resultados estadísticamente fiables, cuando digamos que una tasa de identificación es mejor que otra.

Tabla 13. Número de ficheros identificados, tasa de identificación y banda de fiabilidad de los experimentos realizados con los datos de SES sin normalizar.

	1 Gausiana	2 Gausianas	3 Gausianas	4 Gausianas	5 Gausianas	Nº total de ficheros
EXP 1	183 (87,14% ± 2,22%)	195 (92,86% ± 1,71%)	203 (96,67% ± 1,19%)	198 (94,29% ± 1,54%)	200 (95,24% ± 1,41%)	210
EXP 2	364 (89,66% ± 1,45%)	360 (88,67% ± 1,51%)	390 (96,06% ± 0,93%)	378 (93,1% ± 1,21%)	379 (93,35% ± 1,19%)	406
EXP 3	243 (89,79% ± 1,77%)	234 (86,35% ± 2,00%)	260 (96,06% ± 1,13%)	257 (95,08% ± 1,26%)	251 (92,62% ± 1,52%)	271
EXP 4	50 (89,29% ± 3,97%)	48 (85,71% ± 4,49%)	46 (82,14% ± 4,91%)	50 (89,29% ± 3,97%)	50 (89,29% ± 3,97%)	56
EXP 5	41 (97,62% ± 2,26%)	36 (85,71% ± 5,18%)	41 (97,62% ± 2,26%)	42 (100%)	42 (100%)	42
Mejor resultado	97,62%	92,86%	97,62%	100%	100%	

Podemos observar en la tabla que los mejores resultados se obtienen para el *experimento 5*, que es aquel en el que entrenamos con las frases independientes y las frases del párrafo cuarto y clasificamos los tres primeros párrafos. Se trata de un experimento independiente de texto y, como ya presuponíamos en su descripción, se obtienen tasas de identificación elevadas debido a que la unidad de clasificación es el párrafo (frente a la frase). Adicionalmente, en las frases independientes se supone que habrá más información sobre las emociones, debido a que el locutor tiene mas tiempo para prepararse la forma en la que va a expresar cierta emoción. En cambio, en los párrafos es más complicado interpretar cierta emoción, debido a su mayor duración y a la presencia de pausas. Lo bueno de este experimento es que en el entrenamiento tenemos tanto frases independientes como frases extraídas de párrafos, por lo que los modelos obtenidos son robustos.

Tanto en el *experimento 4* como en el *5* la unidad de clasificación es el párrafo, pero la tasa de identificación obtenida en el *experimento 5* es mucho más elevada. Esta mejora se debe a que los modelos obtenidos en el *experimento 4* se basan en menos datos y asumimos que esto hace que la identificación sea más complicada. Además, en el *experimento 5* empleamos las frases obtenidas del párrafo cuarto en el entrenamiento, lo que hace que en los modelos de cada una de las emociones se consideren tanto las características de frases independientes, como las características de frases extraídas de párrafos. Analizamos a continuación la mejora obtenida en el *experimento 5* frente al *4* en función del número de gausianas empleado. Para ello, restaremos la tasa de identificación media obtenida en el *experimento 5* a la obtenida en el *experimento 4* (es la diferencia entre la tasa de identificación de emociones: ΔE_{acc} (*emotion accuracy*)):

- *1 gausiana:* $\mathbf{\Delta E_{acc}}$ = (97,62% - 89,29%) = **8,33%**
- *2 gausianas:* $\mathbf{\Delta E_{acc}}$ = (85,71% - 85,71%) = **0%**
- *3 gausianas:* $\mathbf{\Delta E_{acc}}$ = (97,62% - 82,14%) = **15,48%**
- *4 gausianas:* $\mathbf{\Delta E_{acc}}$ = (100% - 89,29%) = **10,71%**
- *5 gausianas:* $\mathbf{\Delta E_{acc}}$ = (100% - 89,29%) = **10,71%**

En todos los casos se obtiene una mejora importante (y estadísticamente fiable, como podemos comprobar si observamos las bandas de fiabilidad en la *Tabla 13*), excepto para el caso en el que empleamos 2 gausianas, en el que no mejoramos nada.

El caso en el que utilizamos 2 gausianas es peculiar en todos los experimentos mostrados en la *Tabla 13*, ya que, en general, a medida que aumentamos el número de gausianas utilizadas, aumenta, o al menos se mantiene, la tasa de identificación. Pero al utilizar 2 gausianas, la tasa de identificación disminuye respecto al caso en el que empleamos 1 gausiana. Esta disminución se puede deber al método de división de clusters llevado a cabo en el entrenamiento. Esta peculiaridad sucede en todos los experimentos salvo en el *experimento 1* (en el que entrenamos con párrafos y clasificamos las frases), en el que sí que aumenta la tasa de identificación al pasar de utilizar 1 gausiana a 2. Y es para este experimento, para el que se obtiene la mayor tasa de identificación cuando utilizamos 2 gausianas en el modelo, con un resultado

estadísticamente fiable.

Las tasas de identificación obtenidas en los *experimentos 1, 2 y 3*, que son aquellos en los que la unidad de clasificación es la frase, presentan diferencias no significativas. En estos 3 experimentos se obtiene la mayor tasa de identificación para el caso en el que utilizamos 3 gausianas, y a partir de ahí, la tasa de identificación disminuye cuando usamos 4 y 5 gausianas, pero la caída no es significativa, de hecho las bandas de fiabilidad entre las distintas tasas se solapan.

1.1.1.21. Análisis de la tasa de identificación para cada emoción

A continuación presentamos las matrices de confusión para los experimentos realizados con los vectores sin normalizar obtenidos de los datos de SES. Gracias a estas matrices podremos saber cuales son las emociones que se identifican mejor y cuales las que se identifican peor.

Primero vamos a analizar la media de los resultados de los *experimentos 1, 2* y *3,* en los que la unidad de clasificación es la frase. Y a continuación pasaremos a analizar los resultados de los *experimentos 4* y *5,* en los que la unidad de clasificación es el párrafo. Estos experimentos los analizaremos por separado, dado que son aquellos en los que se obtiene la peor y la mejor tasa de identificación media, respectivamente. En todos los casos, solamente analizaremos los resultados obtenidos cuando utilizamos 1 gausiana, debido a que si analizásemos los resultados en función del distinto número de gausianas empleado, tendríamos demasiados datos y nos resultaría complicado obtener conclusiones.

La siguiente tabla (Tabla 14) muestra la media de la tasa de identificación para cada emoción de los resultados obtenidos utilizando 1 gausiana para los *experimentos 1, 2* y *3.* También se muestra la banda de fiabilidad para las tasas de identificación de la diagonal, es decir, para los casos en los que se identifica cada emoción. En la última fila se muestra la precisión de cada una de las emociones, que se calcula como el cociente entre el número de ficheros de una cierta emoción identificados correctamente y el número total de ficheros que se identifican con dicha emoción (correcta e incorrectamente).

Tabla 14. Media de las tasas de identificación para cada emoción de los *experimentos 1, 2* y *3*, con 1 gausiana.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	80,53% (± 2,77%)	2,31%	15,51%	0,99%	0,66%
Enfado	2,31%	92,08% (± 1,89%)	2,31%		3,30%
Sorpresa	10,23%	0,33%	89,44% (± 2,15%)		
Tristeza		0,66%		93,40% (± 1,73%)	5,94%
Neutro	3,69%			4,15%	92,17% (± 2,25%)
PRECISIÓN	83,23%	96,54%	83,38%	94,79%	90,30%

Observando la tabla llegamos a las siguientes conclusiones:

- La emoción que mejor se identifica es la tristeza (93,4%), seguida de la neutra (92,17%) y el enfado (92,08%). Estas emociones son las que tienen también una precisión más elevada.
- La emoción que peor se identifica es la alegría (80,53%). Su precisión también es la más baja (83,23%), pero muy cercano a la de la sorpresa (83,38%) que se identifica con una tasa mayor (89,44%). Esto se debe a que la alegría se confunde con la sorpresa cuando no se identifica.
- Fijándonos en las bandas de fiabilidad observamos que se solapan las de todas las emociones salvo la de la alegría.
- Hay ciertas emociones que nunca se confunden entre ellas y hace que las precisiones obtenidas sean elevadas:
  - El enfado nunca se confunde con la tristeza.
  - La sorpresa nunca se confunde con la tristeza, ni con la neutra.
  - La tristeza nunca se confunde con la alegría, ni con la sorpresa.
  - La neutra nunca se confunde con el enfado, ni con la sorpresa.

La siguiente tabla (Tabla 15) muestra las tasas de identificación para cada emoción de los resultados obtenidos utilizando 1 gausiana para el *experimento* 4, que es para el que obteníamos las menores tasas de identificación medias. También se muestra

la banda de fiabilidad para las tasas de identificación de la diagonal. Y en la última fila se muestra la precisión de cada una de las emociones.

Tabla 15. Tasas de identificación para cada emoción del *experimento 4,* con 1 gausiana

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	100%				11,1%
Enfado		100%			
Sorpresa	50% (± 13,9%)		50% (± 13,9%)		
Tristeza				100%	
Neutro					100%
PRECISIÓN	67%	100%	100%	100%	90%

En la tabla podemos ver que la alegría, el enfado, la tristeza y la neutra siempre se identifican, mientras que la sorpresa se identifica con la misma tasa con la que se confunde con la alegría (50%). Esta confusión hace que la precisión de la alegría disminuya, obteniéndose un 67%. Y esta confusión también es la que hace que disminuya la tasa de identificación media de este experimento, en el que veíamos que se obtiene la menor de todas ellas.

La siguiente tabla (Tabla 16) muestra las tasas de identificación para cada emoción de los resultados obtenidos utilizando 1 gausiana para el *experimento 5,* que es para el que obteníamos las mayores tasas de identificación medias. También se muestran las bandas de fiabilidad y la precisión de cada una de las emociones.

Tabla 16. Tasas de identificación para cada emoción del *experimento 5*, con 1 gausiana.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	88,9% (± 10,05%)				11,1%
Enfado		100%			
Sorpresa			100%		
Tristeza				100%	
Neutro					100%
PRECISIÓN	100%	100%	100%	100%	90%

En esta tabla observamos que el enfado, la sorpresa, la tristeza y la neutra, siempre se identifican, mientras que la alegría se confunde con la neutra en un 11,1%. A pesar de ello, su precisión es del 100%, es decir, siempre que identificamos un fichero como alegría, lo es. La confusión de la alegría con la neutra hace que la precisión de la neutra sea 90%.

6.1.2. Experimentos de identificación de emociones sobre SES con normalización de características

Una vez analizados los resultados obtenidos en los experimentos de identificación de emociones con vectores de características sin normalizar, el siguiente paso es normalizar estos vectores y evaluar la mejora obtenida.

1.1.1.22. Descripción de los experimentos

Para evaluar la posible mejora obtenida normalizando los vectores de características, realizaremos nuevamente uno de los experimentos realizados con vectores sin normalizar. El experimento seleccionado es aquel en el que entrenábamos con los tres primeros párrafos y clasificábamos las frases independientes y las frases del cuarto párrafo (*experimento* 2). La razón por la que hemos seleccionado este experimento es porque es independiente de texto y presenta una mejor distribución de los datos de entrenamiento y clasificación.

El hecho de que la numeración de los experimentos descritos a continuación comienza en *34* se debe a que a la hora de realizar todos los experimentos con características segmentales, estos fueron los últimos que realizamos.

En función de la normalización empleada (según normalicemos respecto al locutor o respecto a la neutra y si normalizamos respecto a la media y/o a la varianza) obtenemos seis experimentos distintos. En todos ellos, entrenaremos con los vectores normalizados obtenidos de los tres primeros párrafos y clasificaremos los vectores de características normalizados de las frases independientes y las frases del cuarto párrafo.

- EXPERIMENTO 34: *CMN (normalizando respecto a la voz del locutor).*

En este experimento obtendremos la tasa de identificación empleando los vectores de entrenamiento y clasificación normalizados respecto a la media estimada a partir de los datos de la voz del locutor.

La forma de obtener los vectores normalizados es estimando la media de las características de todos los ficheros de entrenamiento y restándosela a cada uno de ellos.

También debemos estimar la media de las características de todos los ficheros de clasificación y restársela a cada uno de ellos.

- EXPERIMENTO 35: *CVN (normalizando respecto a la voz del locutor).*

En este experimento obtendremos la tasa de identificación empleando los vectores de entrenamiento y clasificación normalizados respecto a la varianza estimada a partir de los datos de la voz del locutor.

La forma de obtener los vectores normalizados es estimando la varianza de las características de todos los ficheros de entrenamiento y dividir cada uno de ellos entre la raíz cuadrada de la varianza estimada.

También debemos estimar la varianza de las características de todos los ficheros de clasificación y dividir cada uno de ellos entre la raíz cuadrada de dicha varianza.

- EXPERIMENTO 36: *CMN + CVN (normalizando respecto a la voz del locutor).*

En este experimento obtendremos la tasa de identificación empleando los vectores de entrenamiento y clasificación normalizados respecto a la media y a la varianza estimada a partir de los datos de la voz del locutor.

La forma de obtener los vectores normalizados es estimando la media y la varianza de las características de todos los ficheros de entrenamiento y restando primero la media estimada a cada uno de ellos y a continuación, dividiendo el resultado obtenido entre la raíz cuadrada de la varianza estimada.

También debemos estimar la media y la varianza de las características de todos los ficheros de clasificación, restando primero dicha media y luego dividiendo entre la raíz cuadrada de la varianza estimada, cada uno de ellos.

- EXPERIMENTO 37: *CMN (normalizando respecto a la voz neutra del locutor).*

En este experimento obtendremos la tasa de identificación empleando los vectores de entrenamiento y clasificación normalizados respecto a la media estimada a partir de los datos de la grabación de voz neutra del locutor.

La forma de obtener los vectores normalizados es estimando la media de las características de todos los ficheros de neutra de entrenamiento y restándosela a todos los ficheros de entrenamiento.

También debemos estimar la media de las características de los ficheros de neutra de clasificación y restársela a todos los ficheros de clasificación.

- EXPERIMENTO 38: *CVN (normalizando respecto a la voz neutra del locutor).*

En este experimento obtendremos la tasa de identificación empleando los vectores de entrenamiento y clasificación normalizados respecto a la varianza estimada a partir de los datos de la grabación de voz neutra del locutor.

La forma de obtener los vectores normalizados es estimando la varianza de las características de los ficheros de neutra de entrenamiento y dividir todos los ficheros de entrenamiento entre la raíz cuadrada de la varianza estimada.

También debemos estimar la varianza de las características de los ficheros de neutra de clasificación y dividir todos los ficheros de clasificación entre la raíz cuadrada de la varianza estimada.

- EXPERIMENTO 39: *CMN + CVN (normalizando respecto a la voz neutra del locutor).*

En este experimento obtendremos la tasa de identificación empleando los vectores de entrenamiento y clasificación normalizados respecto a la media y a la varianza estimada a partir de los datos de la grabación de voz neutra del

locutor.

La forma de obtener los vectores normalizados es estimando la media y la varianza de las características de los ficheros de neutra de entrenamiento y restando primero la media estimada a todos los ficheros de entrenamiento y a continuación, dividiendo el resultado obtenido entre la raíz cuadrada de la varianza estimada.

También debemos estimar la media y la varianza de las características de los ficheros de neutra de clasificación, restando primero dicha media y luego dividiendo entre la raíz cuadrada de la varianza estimada, todos los ficheros de clasificación.

1.1.1.23. Resultados de los experimentos

En la Tabla 18 aparece el número total de ficheros identificados, así como la tasa de identificación y la banda de fiabilidad para los seis experimentos descritos en el apartado anterior, realizados con los vectores de características normalizados, obtenidos a partir de los datos de SES. Se muestran los resultados en función del número de gausianas utilizadas en el modelo. En la Tabla 17 se muestran los mismos datos para el mismo tipo de experimento (en el que entrenamos con los tres primeros párrafos y clasificamos las frases independientes y las frases del cuarto párrafo) pero empleando vectores sin normalizar. De esta forma podremos comparar los resultados obtenidos aplicando o prescindiendo de la normalización.

En todos los casos (dado que el experimento realizado siempre es el mismo, lo único que cambia es la normalización o no de los vectores de entrenamiento y clasificación) el número total de ficheros de clasificación es **406**.

Tabla 17. Número de ficheros identificados, tasa de identificación y banda de fiabilidad del *experimento 2* realizado con los datos de SES sin normalizar.

		Sin normalizar
Nº Gausianas	1	364 (89,66% ± 1,45%)
	2	360 (88,67% ± 1,51%)
	3	390 (96,06% ± 0,93%)
	4	378 (93,10% ± 1,21%)
	5	379 (93,35% ± 1,19%)

Tabla 18. Número de ficheros identificados, tasa de identificación y banda de fiabilidad de los *experimentos 34-39,* realizados con los datos de SES normalizados.

		Normalizado respecto a…	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
Nº Gausianas	1	Locutor	366 (90,15% ± 1,42%)	367 (90,39% ± 1,40%)	364 (89,66% ± 1,45%)
	2		356 (87,68% ± 1,57%)	389 (95,81% ± 0,95%)	389 (95,81% ± 0,95%)
	3		390 (96,06% ± 0,93%)	382 (94,09% ± 1,12%)	383 (94,33% ± 1,1%)
	4		379 (93,35% ± 1,18%)	382 (94,09% ± 1,12%)	383 (94,33% ± 1,1%)
	5		373 (91,87% ± 1,30%)	385 (94,83% ± 1,05%)	384 (94,58% ± 1,08%)
	1	Neutra	366 (90,15% ± 1,42%)	367 (90,39% ± 1,40%)	363 (89,41% ± 1,47%)
	2		357 (87,93% ± 1,55%)	390 (96,06% ± 0,93%)	390 (96,06% ± 0,93%)
	3		388 (95,57% ± 0,98%)	382 (94,09% ± 1,12%)	390 (96,06% ± 0,93%)
	4		377 (92,86% ± 1,23%)	380 (93,6% ± 1,17%)	395 (97,29% ± 0,77%)
	5		376 (92,61% ± 1,25%)	383 (94,33% ± 1,1%)	389 (95,81% ± 0,95%)

Podemos observar en los resultados que se muestran en las tablas, que las mejoras obtenidas al normalizar, en general, no son significativas y además no se cumple un patrón en el que siempre se mejore con una cierta normalización o para un cierto número de gausianas empleadas en el modelo.

Analizamos primero los resultados en función del número de gausianas empleadas:

- Para *1 gausiana* los resultados obtenidos con los vectores normalizados o sin normalizar son similares, obteniendo la mayor tasa para el caso en el que normalizamos con la varianza, tanto respecto a la voz del locutor como respecto a su voz neutra. Calculamos la mejora obtenida observando que no es significativa:

  $\mathbf{\Delta E_{acc}}$ = (90,39% - 89,66%) = **0,73%**

- Para *2 gausianas* si que se observan mejoras al normalizar, obteniendo el mejor resultado con dos tipos de normalizaciones diferentes: cuando normalizamos con la varianza y cuando normalizamos con media y varianza, en ambos casos respecto a la voz neutra del locutor.

  $\mathbf{\Delta E_{acc}}$ = (96,06% - 88,67%) = **7,39%**

  En este caso la mejora obtenida si que es significativa, no solapándose las bandas de fiabilidad.

- Para *3 gausianas* no se mejora nada normalizando los vectores de características.

- Para *4 gausianas* se observan mejoras cuando normalizamos los vectores respecto a media y la varianza estimadas a partir de la voz neutra del locutor.

  $\mathbf{\Delta E_{acc}}$ = (97,29% - 93,1%) = **4,19%**

  En este caso la mejora obtenida también es significativa, aunque menor que en el caso en el que utilizamos 2 gausianas.

- Para *5 gausianas*, al igual que en el caso anterior, se mejora normalizando los vectores respecto a la media y la varianza estimadas a partir de la voz neutra del locutor.

  $\mathbf{\Delta E_{acc}}$ = (95,81% - 93,35%) = **2,46%**

  La mejora obtenida en este caso no es muy elevada, pero sus bandas de fiabilidad no se solapan, por lo que podemos decir que se trata de una

mejora estadísticamente fiable (aunque sea pequeña).

Podemos concluir, por tanto, que la normalización con la que se obtienen mejores resultados es aquella en la que normalizamos los vectores respecto a la media y la varianza de la voz neutra del locutor. La tasa de identificación más elevada la obtenemos para el caso en el que utilizamos esta normalización con 4 gausianas (obteniendo un 97.29%). No obstante, la falta de datos hace que, en general, las mejoras obtenidas no sean significativas, solapándose las bandas de fiabilidad.

1.1.1.24. Análisis de la tasa de identificación para cada emoción

A continuación presentamos las matrices de confusión para los experimentos realizados con los vectores normalizados obtenidos sobre los datos de SES. Gracias a estas matrices podremos saber cuales son las emociones que se identifican mejor y cuales las que se identifican peor.

Primero vamos a analizar la media de los resultados de los *experimentos 34, 35* y *36,* en los que normalizamos los vectores respecto a toda la voz del locutor. Y a continuación pasaremos a analizar la media de los resultados de los *experimentos 37, 38* y *39,* en los que normalizamos los vectores respecto a la grabación de voz neutra del locutor. En ambos casos, analizaremos los resultados obtenidos utilizando 1, 2, 3, 4 y 5 gausianas.

Las siguientes tablas (Tabla 19-*Tabla 23*) muestran la media de la tasa de identificación para cada emoción de los resultados obtenidos utilizando 1, 2, 3, 4 y 5 gausianas, respectivamente, para los *experimentos 34, 35* y *36.* También se muestra la banda de fiabilidad para las tasas de identificación de la diagonal. En la última fila se muestra la precisión de cada una de las emociones.

Tabla 19. Media de las tasas de identificación para cada emoción de los *experimentos 34, 35 y 36*, utilizando 1 gausiana.

1 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	83,14% (± 2,22%)	3,07%	10,73%	2,68%	0,38%
Enfado	1,92%	91,95% (± 1,62%)	1,92%		4,21%
Sorpresa	5,36%	2,68%	91,19% (± 1,68%)	0,77%	
Tristeza		1,53%		93,49% (± 1,47%)	4,98%
Neutro	4,02%	0,57%		4,60%	90,80%(± 2,10%)
PRECISIÓN	88,03%	92,13%	87,82%	92,08%	90,46%

Tabla 20. Media de las tasas de identificación para cada emoción de los *experimentos 34, 35 y 36*, utilizando 2 gausianas.

2 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	90,42% (± 1,75%)	1,15%	3,83%	3,07%	1,53%
Enfado	1,53%	96,93% (± 1,02%)			1,53%
Sorpresa	13,41%	4,21%	81,23% (± 2,32%)	0,77%	0,38%
Tristeza				100%	
Neutro				1,15%	98,85% (± 0,78%)
PRECISIÓN	85,82%	94,76%	95,50%	95,26%	96,63%

Tabla 21. Media de las tasas de identificación para cada emoción de los *experimentos 34, 35 y 36*, utilizando 3 gausianas.

3 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	90,80% (± 1,72%)	4,21%	3,83%	0,77%	0,38%
Enfado	0,38%	98,85% (± 0,63%)	0,38%		0,38%
Sorpresa	3,07%	3,83%	93,10% (± 1,51%)		
Tristeza		2,68%		97,32% (± 0,96%)	
Neutro		6,32%			93,68%(± 1,77%)
PRECISIÓN	96,34%	85,29%	95,67%	99,22%	99,19%

Tabla 22. Media de las tasas de identificación para cada emoción de los *experimentos 34, 35 y 36*, utilizando 4 gausianas.

4 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	86,97% (± 2%)	5,75%	6,13%	0,77%	0,38%
Enfado	0,38%	98,85% (± 0,63%)	0,38%		0,38%
Sorpresa	4,21%	4,98%	90,80% (± 1,72%)		
Tristeza		1,15%		98,08% (± 0,81%)	0,77%
Neutro		4,60%			95,40% (± 1,52%)
PRECISIÓN	94,98%	85,71%	93,31%	99,22%	98,42%

Tabla 23. Media de las tasas de identificación para cada emoción de los *experimentos 34, 35 y 36*, utilizando 5 gausianas.

5 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	86,97% (± 2%)	4,60%	7,28%	0,77%	0,38%
Enfado		99,23% (± 0,52%)			0,77%
Sorpresa	5,75%	4,98%	89,27% (± 1,84%)		
Tristeza		0,77%		98,47% (± 0,73%)	0,77%
Neutro		4,60%			95,40% (± 1,52%)
PRECISIÓN	93,80%	86,91%	92,46%	99,23%	98,03%

Las conclusiones que sacamos al examinar estas tablas son:

- Las emociones que mejor se identifican son la tristeza y el enfado. La mayor tasa de identificación (100%) se obtiene para la tristeza en el caso en el que utilizamos 2 gausianas (Tabla 20). La precisión más elevada (99,23%) también se obtiene para la tristeza, pero para el caso en el que utilizamos 5 gausianas (Tabla 23).
- Las emociones que peor se identifican son la alegría y la sorpresa. Para los casos en los que utilizamos 2 y 5 gausianas, sus bandas de fiabilidad se solapan, por lo que aunque en ambos casos, la tasa de identificación de la sorpresa es ligeramente superior a la de la alegría, los datos no son estadísticamente fiables. La peor tasa de identificación (81,23%) se obtiene

para la sorpresa, para el caso en el que utilizamos 2 gausianas (Tabla 20).

- A pesar de que el enfado tiene una tasa de identificación elevada, es la emoción con menor precisión, cuando utilizamos 3, 4 y 5 gausianas, ya que en estos casos todas las emociones se identifican alguna vez como enfado. La menor de todas ellas (85,29%) se obtiene cuando utilizamos 3 gausianas (Tabla 21).
- Cuando utilizamos 1 gausiana, la alegría se confunde con la sorpresa (10,73% -Tabla 19). Y cuando utilizamos 2 gausianas, la sorpresa se confunde con la alegría (13,41% - Tabla 20).
- Las emociones que nunca se confunden con otras son:
  - El enfado nunca se confunde con la tristeza.
  - La tristeza nunca se confunde con la alegría ni con la sorpresa.
  - La neutra nunca se confunde con la sorpresa y sólo se confunde con la alegría cuando utilizamos 1 gausiana (Tabla 19*)*.
  - La sorpresa sólo se confunde con el estado neutro cuando utilizamos 2 gausianas *(Tabla 20*).

En la siguiente tabla (Tabla 24) se muestra la precisión media para los *experimentos 34, 35 y 36*, en función del número de gausianas. La razón por la que extraemos en una tabla la precisión que ya veíamos en las tablas anteriores (Tabla 19-*Tabla 23*), es para evaluar como influye el número de gausianas en la precisión de las distintas emociones.

Tabla 24. Precisión de cada emoción para la media de los *experimentos 34, 35* y *36*.

	Alegría	Enfado	Sorpresa	Tristeza	Neutro
1 GAUS	88,03%	92,13%	87,82%	92,08%	90,46%
2 GAUS	85,82%	94,76%	95,50%	95,26%	96,63%
3 GAUS	96,34%	85,29%	95,67%	99,22%	99,19%
4 GAUS	94,98%	85,71%	93,31%	99,22%	98,42%
5 GAUS	93,80%	86,91%	92,46%	99,23%	98,03%

En general, las precisiones más elevadas se consiguen cuando utilizamos 3 gausianas en nuestro modelo. El enfado es la única emoción para la que no se cumple,

obteniéndose la mayor precisión cuando utilizamos 2 gausianas.

Para la sorpresa y la neutra, se cumple que la precisión aumenta en función del número de gausianas hasta que se llega a 3 gausianas, y a partir de ahí, disminuye. Para la tristeza también se cumple esto, salvo que para 4 y 5 gausianas se mantiene la precisión obtenida al utilizar 3 gausianas.

A continuación analizamos los resultados de los experimentos realizados con los vectores normalizados respecto a la voz neutra del locutor. Las siguientes tablas (Tabla 25-*Tabla 29*) muestran la media de la tasa de identificación para cada emoción de los resultados obtenidos utilizando 1, 2, 3, 4 y 5 gausianas, respectivamente, para los *experimentos 37, 38* y *39.* También se muestra la banda de fiabilidad para las tasas de identificación de la diagonal, es decir, para los casos en los que se identifica cada emoción. En la última fila se muestra la precisión de cada una de las emociones.

Tabla 25. Media de las tasas de identificación para cada emoción de los *experimentos 37, 38 y 39*, utilizando 1 gausiana.

1 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	81,23% (± 2,32%)	2,68%	13,03%	2,68%	0,38%
Enfado	2,30%	91,19% (± 1,68%)	2,30%		4,21%
Sorpresa	3,83%	2,68%	92,72% (± 1,54%)	0,77%	
Tristeza		1,53%		93,87% (± 1,43%)	4,60%
Neutro	4,02%			4,60%	91,38% (± 2,04%)
PRECISIÓN	88,89%	92,97%	85,82%	92,11%	90,86%

Tabla 26. Media de las tasas de identificación para cada emoción de los *experimentos 37, 38 y 39*, utilizando 2 gausianas.

2 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	90,80% (± 1,72%)	1,15%	4,21%	2,68%	1,15%
Enfado	1,92%	96,55% (± 1,08%)			1,53%
Sorpresa	11,49%	4,60%	83,14% (± 2,2%)	0,77%	
Tristeza		0,77%		98,85% (± 0,63%)	0,38%
Neutro				0,57%	99,43% (± 0,55%)
PRECISIÓN	87,13%	93,68%	95,18%	96,09%	97,01%

Tabla 27. Media de las tasas de identificación para cada emoción de los *experimentos 37, 38 y 39*, utilizando 3 gausianas.

3 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	87,74% (± 1,95%)	4,98%	4,98%	1,15%	1,15%
Enfado	0,38%	98,47% (± 0,73%)	0,38%		0,77%
Sorpresa	1,53%	4,60%	93,49% (± 1,43%)	0,38%	
Tristeza		0,38%		98,08% (± 0,81%)	1,53%
Neutro					100%
PRECISIÓN	97,86%	90,81%	94,57%	98,46%	96,67%

Tabla 28. Media de las tasas de identificación para cada emoción de los *experimentos 37, 38 y 39*, utilizando 4 gausianas.

4 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	85,82% (± 2,07%)	7,28%	4,98%	1,53%	0,38%
Enfado	0,38%	98,08% (± 0,81%)	0,77%		0,77%
Sorpresa	1,53%	4,21%	94,25% (± 1,38%)		
Tristeza		0,38%		98,85% (0,63%)	0,77%
Neutro	1,15%	1,72%		0,57%	96,55% (1,33%)
PRECISIÓN	96,55%	87,82%	94,25%	97,91%	98,05%

Tabla 29. Media de las tasas de identificación para cada emoción de los *experimentos 37, 38 y 39*, utilizando 5 gausianas.

5 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	82,76% (± 2,24%)	6,51%	9,58%	0,77%	0,38%
Enfado		100%			
Sorpresa	1,53%	4,98%	93,10% (± 1,51%)	0,38%	
Tristeza		0,38%		98,85% (± 0,63%)	0,77%
Neutro	0,57%	1,15%		0,57%	97,70% (± 1,09%)
PRECISIÓN	97,52%	88,47%	90,67%	98,29%	98,84%

Las conclusiones que sacamos al examinar estas tablas son:

- Las emociones que mejor se identifican son el enfado, la tristeza y la neutra. La neutra siempre se identifica cuando utilizamos 3 gausianas (Tabla 27) y el enfado siempre se identifica cuando utilizamos 5 gausianas (Tabla 29).
- Las precisiones más elevadas también se obtienen para el enfado, la tristeza y, principalmente, para la neutra. La mayor precisión (98,84%) la obtiene la voz neutra en el caso en el que utilizamos 5 gausianas (Tabla 29).
- Las emociones que peor se identifican son la alegría y la sorpresa, aunque no para todos los casos. Lo analizamos en función del número de gausianas empleado:
    - *1 gausiana (Tabla 25):* la alegría se confunde con la sorpresa (13,03%), obteniéndose en este caso la menor tasa de identificación (81,23%). En cambio, la sorpresa tiene una elevada tasa de identificación (92,72%).
    - *2 gausianas (Tabla 26):* la sorpresa se confunde con la alegría (11,49%), y la alegría se confunde con todas las demás emociones, pero no con una elevada tasa de confusión (con la que más se confunde es con la sorpresa, con un 4,21%).

- *3 gausianas (Tabla 27):* la menor tasa de identificación la tiene la alegría (87,74%), pero su precisión en este caso es elevada (97,86%).
- *4 gausianas (Tabla 28):* la alegría se confunde principalmente con el enfado (7,28%) y también con la sorpresa (4,98%). La tasa de identificación de la sorpresa es elevada en este caso (94,25%), solapándose la banda de fiabilidad con la de la neutra.
- *5 gausianas (Tabla 29):* la tasa de identificación de la alegría disminuye (82,76%), confundiéndose con la sorpresa (9,58%) y con el enfado (6,51%). Pero su precisión, al igual que cuando utilizábamos 3 gausianas, también es elevada (97,52%).

- A pesar de que el enfado tiene una tasa de identificación elevada, es la emoción con menor precisión, cuando utilizamos 3, 4 y 5 gausianas. Pero la menor precisión (85,82%) se obtiene para la sorpresa en el caso en el que utilizamos 1 gausiana *(Tabla 25)*.
- Al igual que cuando normalizábamos respecto a la voz del locutor, el enfado nunca se confunde con la tristeza, la tristeza nunca se confunde con la alegría ni con la sorpresa, la neutra nunca se confunde con la sorpresa y la sorpresa nunca se confunde con la neutra.

En la Tabla 30 se muestra la precisión media para los *experimentos 37, 38 y 39*, en función del número de gausianas, para evaluar la influencia del número de gausianas en la precisión de las distintas emociones.

Tabla 30. Precisión de cada emoción para la media de los *experimentos 37, 38* y *39*.

	Alegría	Enfado	Sorpresa	Tristeza	Neutro
1 GAUS	88,89%	92,97%	85,82%	92,11%	90,86%
2 GAUS	87,13%	93,68%	95,18%	96,09%	97,01%
3 GAUS	97,86%	90,81%	94,57%	98,46%	96,67%
4 GAUS	96,55%	87,82%	94,25%	97,91%	98,05%
5 GAUS	97,52%	88,47%	90,67%	98,29%	98,84%

Observando esta tabla podemos comprobar que el número de gausianas influye de distinta forma en la precisión de cada una de las emociones, así que las analizaremos por separado:

- *Alegría:* aumenta la precisión al aumentar el número de gausianas hasta 3, pero a partir de aquí, se mantiene prácticamente invariable.
- *Enfado:* en general, disminuye la precisión al aumentar el número de gausianas utilizado.
- *Sorpresa:* aumenta la precisión al pasar de utilizar 1 a 2 gausianas, pero a partir de ahí disminuye.
- *Tristeza:* al igual que pasaba con la alegría, aumenta la precisión al aumentar el número de gausianas hasta 3, y a partir de ahí, más o menos se mantiene.
- *Neutra:* en general, aumenta la precisión al aumentar el número de gausianas utilizado.

Por tanto, concluimos que salvo para la neutra, el uso de 4 y 5 gausianas no hace que aumentemos el grado de precisión.

Es interesante comprobar a que emoción le sienta mejor la normalización, es decir, comprobar cual es la emoción en la que conseguimos un mayor aumento de la tasa de identificación al emplear vectores normalizados. Para ello, en las siguientes tablas (Tabla 31-*Tabla 35*) se muestran las tasas de identificación para cada emoción para el *experimento 2*, en el que utilizábamos vectores sin normalizar, para 1, 2, 3, 4 y 5 gausianas:

Tabla 31. Tasas de identificación para cada emoción del *experimento 2,* utilizando 1 gausiana.

1 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	80,46% (± 4,08%)	2,30%	14,94%	1,15%	1,15%
Enfado	2,30%	90,80% (± 2,97%)	2,30%		4,60%
Sorpresa	6,90%	1,15%	91,95% (±2,8%)		
Tristeza		1,15%		91,95% (± 2,8%)	6,90%
Neutro	3,45%			1,72%	94,83% (± 2,79%)
PRECISIÓN	86,42%	95,18%	84,21%	96,97%	88,24%

Tabla 32. Tasas de identificación para cada emoción del *experimento 2,* utilizando 2 gausianas.

2 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	93,10% (± 2,61%)	1,15%	1,15%	3,45%	1,15%
Enfado	1,15%	96,55% (± 1,88%)			2,30%
Sorpresa	27,59%	11,49%	59,77% (± 5,05%)	1,15%	
Tristeza		1,15%		98,85% (± 1,1%)	
Neutro				1,72%	98,28% (± 1,64%)
PRECISIÓN	76,42%	87,50%	98,11%	93,99%	96,61%

Tabla 33. Tasas de identificación para cada emoción del *experimento 2,* utilizando 3 gausianas.

3 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	88,51% (± 3,28%)	2,30%	8,05%		1,15%
Enfado	1,15%	96,55% (± 1,88%)	1,15%		1,15%
Sorpresa	2,30%		97,70% (±1,54%)		
Tristeza		1,15%		98,85% (± 1,1%)	
Neutro					100%
PRECISIÓN	96,25%	96,55%	91,40%	100%	97,75%

Tabla 34. Tasas de identificación para cada emoción del *experimento 2,* utilizando 4 gausianas.

4 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	79,31% (± 4,17%)	10,34%	8,05%	1,15%	1,15%
Enfado	1,15%	95,40% (± 2,16%)	2,30%		1,15%
Sorpresa	3,45%	2,30%	94,25% (± 2,4%)		
Tristeza				100%	
Neutro		1,72%			98,28% (± 1,64%)
PRECISIÓN	94,52%	86,91%	90,11%	98,86%	97,71%

Tabla 35. Tasas de identificación para cada emoción del *experimento 2,* utilizando 5 gausianas.

5 GAUS	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	80,46% (± 4,08%)	11,49%	6,90%		1,15%
Enfado		98,85% (± 1,1%)			1,15%
Sorpresa	3,45%	5,75%	90,80% (± 2,97%)		
Tristeza				100%	
Neutro		1,72%			98,28% (± 1,64%)
PRECISIÓN	95,89%	83,90%	92,94%	100%	97,71%

Comparando los resultados obtenidos en estas tablas (Tabla 31-*Tabla 35*) con los que obteníamos en las tablas normalizando respecto a la voz locutor (Tabla 19-*Tabla 23*) y en las tablas normalizando respecto a la voz neutra (Tabla 25-*Tabla 29*), obtenemos las siguientes conclusiones para cada una de las emociones:

- *Alegría:* sólo conseguimos mejoras significativas cuando normalizamos, tanto respecto al locutor como a la neutra, utilizando 4 gausianas; y cuando normalizamos respecto a la voz del locutor utilizando 5 gausianas. La mayor mejora obtenida es la siguiente, cuando normalizamos respecto a la voz del locutor utilizando 4 gausianas:

  $\mathbf{\Delta E_{acc}}$ = (86,97% - 79,31%) = **7,66%**

- *Enfado:* sólo conseguimos mejoras significativas cuando normalizamos tanto respecto a la voz del locutor como a su voz neutra, utilizando 4 gausianas; y cuando normalizamos respecto a la voz neutra, utilizando 5 gausianas. La mayor mejora conseguida es la siguiente, cuando normalizamos respecto a la voz del locutor utilizando 4 gausianas:

  $\mathbf{\Delta E_{acc}}$ = (98,85% - 95,4%) = **3,45%**

- *Sorpresa:* conseguimos aumentar la tasa de identificación al normalizar, tanto respecto al locutor como a la neutra, sólo en los casos en los que utilizamos 2 y 3 gausianas. La mayor mejora se obtiene cuando normalizamos respecto a la

voz neutra utilizando 2 gausianas. Esta mejora es elevada, ya que la tasa de identificación obtenida sin normalizar es bastante baja:

$\mathbf{\Delta E_{acc}}$ = (83,14% - 59,77%) = **23,37%**

- *Tristeza:* las tasas de identificación obtenidas para la tristeza son elevadas, y no conseguimos mejorarlas al utilizar vectores normalizados, salvo en el caso de emplear 2 gausianas normalizando respecto a la voz del locutor, que obtenemos la siguiente mejora:

  $\mathbf{\Delta E_{acc}}$ = (100% - 98,85%) = **1,15%**

  Cuando utilizamos 4 y 5 gausianas, conseguimos identificar siempre está emoción (100% de tasa de identificación) al utilizar vectores sin normalizar. En cambio, esta tasa disminuye al normalizarlos.

- *Neutra:* en ninguno de los casos conseguimos mejorar la tasa de identificación al normalizar los vectores de características.

Dadas las diferencias entre las tasas de identificación de las distintas emociones resulta conveniente analizar la mejora relativa del error, en lugar de la simplemente la mejora absoluta, que es la que hemos calculado. La mejora relativa del error se define como el cociente entre la mejora absoluta del error (que es igual a la mejora absoluta que hemos calculado) entre la tasa de error obtenida sin normalizar. Aplicándolo a cada una de las emociones, salvo a la neutra, cuya tasa no sufre ninguna mejora al normalizar, obtenemos lo siguiente:

- *Alegría:* (86,97% - 79,31%) / (100% - 79,31%) = **37,02%**
- *Enfado:* (98,85% - 95,4%) / (100% - 95,4%) = **75%**
- *Sorpresa:* (83,14% - 59,77%) / (100% - 59,77%) = **58,09%**
- *Tristeza:* (100% - 98,85%) / (100% - 98,85%) = **1%**

Fijándonos en la mejora relativa del error, a la emoción que mejor le sienta la normalización es al enfado; mientras que fijándonos en la mejora absoluta, a la emoción que mejor le sienta la normalización es a la sorpresa.

A continuación analizamos la influencia de la normalización en la precisión, en función del número de gausianas empleado:

- *1 gausiana:* aumenta ligeramente la precisión de la alegría, sorpresa y neutra; y disminuye la del enfado y la tristeza.
- *2 gausianas:* aumenta la precisión de la alegría, el enfado y la tristeza; se mantiene más o menos la de la neutra; y disminuye la de la sorpresa.
- *3 gausianas:* aumenta la precisión de la alegría y la sorpresa; y disminuye la del enfado y la tristeza. La de la neutra aumenta cuando normalizamos respecto a la voz del locutor.
- *4 gausianas:* aumenta la precisión de la alegría (ligeramente), de la sorpresa y de la neutra; y se mantiene más o menos la del enfado.
- *5 gausianas:* aumenta la precisión del enfado y de la neutra; y disminuye el de la tristeza. La precisión de la alegría y la sorpresa, en unos casos aumenta y en otros disminuye.

En general, podemos decir que al normalizar aumenta la precisión de la alegría, la sorpresa y la neutra; y disminuye la del enfado y la tristeza; aunque como hemos analizado detalladamente, esto no se cumple para todas las gausianas.

6.1.3. Conclusiones de los experimentos de identificación de emociones sobre SES

Una vez analizados los resultados obtenidos en los distintos experimentos de identificación realizados con la base de datos SES con características basadas en MFCC, las conclusiones a las que llegamos son las siguientes:

- De los experimentos realizados con vectores de características sin normalizar, con el que obtenemos mejor resultado es aquel en el que entrenamos con frases (tanto las frases independientes, como las frases extraídas del cuarto párrafo) y clasificamos los tres primeros párrafos de cada bloque. Al entrenar con frases de dos tipos (independientes y extraídas de párrafos) obtenemos unos modelos robustos, que hacen que obtengamos tan buenos resultados a la hora de clasificar. También influye

que al ser la unidad de clasificación el párrafo, tengamos un mayor número de vectores de clasificación, lo que hace que sea más fácil que a lo largo del párrafo se identifiquen los distintos vectores con la emoción correcta.

- El experimento en el que entrenamos con las frases independientes y clasificamos los párrafos, es en el que obtenemos peores resultados, a pesar de que también tenemos el párrafo como unidad de clasificación, pero en este caso el número de vectores de entrenamiento es mucho menor que el de vectores de clasificación y, por tanto, los modelos obtenidos se basan en pocos datos.

- Los resultados obtenidos en los tres experimentos en los que la unidad de clasificación es la frase son similares. Las diferencias obtenidas entre los experimentos dependientes de texto y el que no lo es, no son significativas.

- En general, las tasas de identificación medias aumentan al aumentar el número de gausianas, excepto cuando empleamos 2 gausianas.

- Las emociones que mejor se identifican en todos los casos son la tristeza, el enfado y la neutra, consiguiendo identificarlas siempre en los siguientes casos:

  - *Tristeza:* cuando utilizamos el párrafo como unidad de clasificación (con 1 gausiana) (*experimentos 4 y 5)* y cuando normalizamos respecto a la voz del locutor (con 2 gausianas).

  - *Enfado y neutra:* cuando utilizamos el párrafo como unidad de clasificación (con 1 gausiana) (*experimentos 4 y 5)* y cuando normalizamos respecto a la voz neutra (con 5 gausianas el enfado y con 3 gausianas la neutra).

- Las emociones que peor se identifican son la alegría, seguida de la sorpresa. En general se suelen confundir la una con la otra. Es importante resaltar que se trata de las dos únicas emociones positivas disponibles en SES. Sin embargo, en algunos experimentos conseguimos identificarlas siempre: la alegría en el *experimento 4* y la sorpresa en el *experimento 5.*

- En general, la normalización con la que obtenemos mejores resultados es aquella en la que normalizamos respecto a la media y a la varianza de la voz neutra del locutor, aunque esto no se cumple para todas las gausianas. Los mejores resultados al normalizar los obtenemos cuando empleamos 2 gausianas, obteniendo una mejora de 7,39%. Pero el caso en el que utilizamos 2 gausianas es especial, ya que obtenemos resultados no previsibles. El siguiente mejor resultado al normalizar es cuando utilizamos 4 gausianas, obteniendo una mejora de 4,19%.

- En todos los experimentos realizados en este apartado se cumple que:
  - El enfado nunca se confunde con la tristeza.
  - La tristeza nunca se confunde con la alegría, ni con la sorpresa.
  - La sorpresa y la neutra nunca se confunden entre ellas.

- La emoción con la que conseguimos una mayor mejora absoluta al normalizar los vectores de características es la sorpresa, consiguiendo una mejora de un 23,37%, debido a que la tasa obtenida con vectores sin normalizar es muy baja (59,77%). Sin embargo, fijándonos en la mejora relativa del error, es el enfado el que obtiene el mayor valor.

- Después de la mejora de la sorpresa, las mejoras más significativas se obtienen para la alegría y el enfado, normalizando respecto a la voz del locutor, con 4 gausianas.

- En general, normalizando conseguimos aumentar la precisión de la alegría, la sorpresa y la neutra; pero disminuye la del enfado y la tristeza. El enfado es el que tiene la menor precisión en los experimentos en los que empleamos vectores normalizados.

6.2.Identificación con EMODB

En este apartado describimos y analizamos los resultados obtenidos a partir de los distintos experimentos de identificación de emociones sobre EMODB basados en MFCC.

6.2.1.Descripción de los experimentos

En EMODB disponemos de ficheros de 10 locutores distintos, de forma que en cada uno de los experimentos realizados, entrenaremos con 9 de los locutores y clasificaremos el que nos queda, e iremos rotando el locutor utilizado en el entrenamiento, obteniendo un total de 10 experimentos dentro de cada uno de los experimentos descritos.

El número de vectores de entrenamiento y clasificación dependerá del locutor que usemos para la clasificación. Pero en todos los casos se cumple que el número de vectores de entrenamiento es un orden de magnitud mayor que el número de vectores de clasificación. Más o menos en media tenemos 130.000 (90% de todos los datos) vectores de entrenamiento, frente a 14.000 (10% de todos los datos) vectores de clasificación.

A continuación describimos cada uno de los experimentos realizados con los datos de EMODB, en los que evaluaremos las distintas técnicas de normalización que fueron explicadas en el *apartado 1.1.1.17*:

- EXPERIMENTO 6: *Sin normalizar los vectores de características.*

  En la realización de estos experimentos empleamos los vectores de características obtenidos a partir de los datos de EMODB sin normalizar. Haremos diversos experimentos en los que entrenemos con los vectores de características de 9 de los locutores y clasifiquemos con el que nos queda, e iremos rotando el locutor que usamos en la clasificación.

- EXPERIMENTO 7: *CMN (normalizando respecto a la voz del locutor).*

  En este experimento obtendremos la tasa de identificación empleando vectores de entrenamiento y clasificación normalizados respecto a la media estimada a partir de los datos de cada locutor.

La forma de obtener los vectores normalizados es estimando la media de las características de todos los ficheros de cada uno de los locutores de entrenamiento y restando dicha media a cada uno de las características de sus ficheros.

También debemos estimar la media de las características de todos los ficheros del locutor que vamos a clasificar y restársela a cada uno de ellos.

- EXPERIMENTO 8: *CVN (normalizando respecto a la voz del locutor).*

  En este experimento obtendremos la tasa de identificación empleando vectores de entrenamiento y clasificación normalizados respecto a la varianza estimada a partir de los datos de cada locutor.

  La forma de obtener los vectores normalizados es estimando la varianza de las características de todos los ficheros de cada uno de los locutores de entrenamiento y dividiendo las características de los ficheros de cada locutor entre la raíz cuadrada de la varianza estimada para dicho locutor.

  También debemos estimar la varianza de las características de todos los ficheros del locutor que vamos a clasificar y dividir cada uno de ellos entre la raíz cuadrada de la varianza.

- EXPERIMENTO 9: *CMN + CVN (normalizando respecto a la voz del locutor).*

  En este experimento obtendremos la tasa de identificación empleando vectores de entrenamiento y clasificación normalizados respecto a la media y la varianza estimada a partir de los datos de cada locutor.

  La forma de obtener los vectores normalizados es estimando la media y la varianza de las características de todos los ficheros de cada uno de los locutores de entrenamiento y restando primero la media obtenida para cada locutor a cada uno de las características de sus ficheros y, a continuación, dividiendo este resultado entre la raíz cuadrada de la varianza estimada.

  También debemos estimar la media y la varianza de las características de todos los ficheros del locutor que vamos a clasificar, restando primero la media a cada uno de ellos, y dividiendo el resultado obtenido entre la raíz cuadrada

de la varianza estimada.

- EXPERIMENTO 10: *CMN (normalizando respecto a la voz neutra del locutor).*

  En este experimento obtendremos la tasa de identificación empleando vectores de entrenamiento y clasificación normalizados respecto a la media estimada a partir de los datos de las grabaciones de voz neutra de cada locutor.

  La forma de obtener los vectores normalizados es estimando la media de las características de los ficheros de voz neutra de cada locutor de entrenamiento y restando la media obtenida para cada locutor a cada uno de las características de todos sus ficheros.

  También debemos estimar la media de las características de los ficheros de voz neutra del locutor que vamos a clasificar y restársela a cada uno de las características de todos los ficheros del locutor.

- EXPERIMENTO 11: *CVN (normalizando respecto a la voz neutra del locutor).*

  En este experimento obtendremos la tasa de identificación empleando vectores de entrenamiento y clasificación normalizados respecto a la varianza estimada a partir de los datos de las grabaciones de voz neutra de cada locutor.

  La forma de obtener los vectores normalizados es estimando la varianza de las características de los ficheros de voz neutra de cada locutor de entrenamiento y dividiendo las características de todos los ficheros de cada locutor entre la raíz cuadrada de la varianza estimada para dicho locutor.

  También debemos estimar la varianza de las características de los ficheros de neutra del locutor que vamos a clasificar y dividir cada uno de las características de todos los ficheros entre la raíz cuadrada de dicha varianza.

- EXPERIMENTO 12: *CMN + CVN (normalizando respecto a la voz neutra del locutor).*

  En este experimento obtendremos la tasa de identificación empleando vectores de entrenamiento y clasificación normalizados respecto a la media y la varianza estimada a partir de los datos de las grabaciones de voz neutra de cada locutor.

La forma de obtener los vectores normalizados es estimando la media y la varianza de las características de los ficheros de voz neutra de cada locutor de entrenamiento y restando primero la media obtenida para cada locutor a cada una de las características de todos sus ficheros y, a continuación, dividiendo este resultado entre la raíz cuadrada de la varianza estimada.

También debemos estimar la media y la varianza de las características de los ficheros de voz neutra del locutor que vamos a clasificar, restando primero la media a cada uno de las características de todos los ficheros del locutor, y dividiendo el resultado obtenido entre la raíz cuadrada de la varianza estimada.

6.2.2.Resultados de los experimentos

A continuación, mostramos en las Tabla 36-*Tabla 45,* el número de ficheros identificados correctamente, así como la tasa de identificación y la banda de fiabilidad para cada uno de los experimentos explicados en el apartado anterior. Los resultados se muestran en función del locutor destinado a la clasificación.

Para poder estudiar el efecto de la normalización, se muestran en tablas diferentes los resultados obtenidos para el *experimento 6*, en el que se utilizan vectores de características sin normalizar, y los resultados obtenidos para el resto de experimentos explicados anteriormente (*experimentos 7, 8, 9, 10, 11* y *12*), en los que se utilizan vectores de características normalizados. Las distintas tablas presentadas se diferencian en el número de gausianas utilizadas, que va de 1 a 5.

Tabla 36. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad del *experimento 6,* realizado con los datos de EMODB sin normalizar, para cada uno de los locutores, para 1 gausiana.

1 GAUS		Sin normalizar
N° Locutor	3	28/49 (57,14% ± 6,79%)
	8	25/58 (43,1% ± 6,24%)
	9	20/43 (46,51% ± 7,3%)
	10	5/38 (13,16% ± 5,26%)
	11	21/55 (38,18% ± 5,26%)
	12	13/35 (37,14% ± 7,84%)
	13	39/61 (63,93% ± 5,90%)
	14	43/69 (62,32% ± 5,60%)
	15	15/56 (26,79% ± 5,68%)
	16	24/71 (33,8% ± 5,39%)

Tabla 37. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad de los *experimentos 7-12,* realizados con los datos de EMODB normalizados, para cada uno de los locutores, para 1 gausiana.

1 GAUS		*Normalizado respecto a…*	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Locutor	3	Locutor	29/49 (59,18% ± 6,74%)	24/49 (48,98% ± 6,86%)	26/49 (53,06% ± 6,84%)
	8		26/58 (44,83% ± 6,27%)	23/58 (39,66% ± 6,17%)	28/58 (48,28% ± 6,3%)
	9		27/43 (62,79% ± 7,08%)	21/43 (48,84% ± 7,32%)	19/43 (44,19% ± 7,27%)
	10		15/38 (39,47% ± 7,61%)	13/38 (34,21% ± 7,39%)	19/38 (50% ± 7,79%)
	11		32/55 (58,18% ± 6,39%)	27/55 (49,09% ± 6,47%)	31/55 (56,36% ± 6,42%)
	12		17/35 (48,57% ± 8,11%)	15/35 (42,86% ± 8,03%)	15/35 (42,86% ± 8,03%)
	13		42/61 (68,85% ± 5,69%)	38/61 (62,3% ± 5,96%)	40/61 (65,57% ± 5,84%)
	14		51/69 (73,91% ± 5,07%)	47/69 (68,12% ± 5,39%)	51/69 (73,91% ± 5,07%)
	15		29/56 (51,79% ± 6,41%)	22/56 (39,29% ± 6,27%)	32/56 (57,14% ± 6,35%)
	16		44/71 (61,97% ± 5,53%)	30/71 (42,25% ± 5,63%)	53/71 (74,65% ± 4,96%)
N° Locutor	3	Neutra	28/49 (57,14% ± 6,79%)	26/49 (53,06% ± 6,84%)	26/49 (53,06% ± 6,84%)
	8		31/58 (53,45% ± 6,29%)	25/58 (43,1% ± 6,24%)	32/58 (55,17% ± 6,27%)
	9		25/43 (58,14% ± 7,22%)	20/43 (46,51% ± 7,3%)	29/43 (67,44% ± 6,86%)
	10		18/38 (47,37% ± 7,78%)	9/38 (23,68% ± 6,62%)	16/38 (42,11% ± 7,69%)
	11		29/55 (52,73% ± 6,46%)	26/55 (47,27% ± 6,46%)	30/55 (54,55% ± 6,45%)
	12		14/35 (40% ± 7,95%)	15/35 (42,86% ± 8,03%)	13/35 (37,14% ± 7,84%)
	13		38/61 (62,3% ± 5,69%)	41/61 (67,21% ± 5,77%)	40/61 (65,57% ± 5,84%)
	14		51/69 (73,91% ± 5,07%)	49/69 (71,01% ± 5,24%)	48/69 (69,57% ± 5,32%)
	15		29/56 (51,79% ± 6,41%)	23/56 (41,07% ± 6,31%)	29/56 (51,79% ± 6,41%)
	16		45/71 (63,38% ± 5,49%)	27/71 (38,03% ± 5,53%)	51/71 (71,83% ± 5,12%)

Tabla 38. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad del *experimento 6,* realizado con los datos de EMODB sin normalizar, para cada

uno de los locutores, para 2 gausianas.

2 GAUS		Sin normalizar
N° Locutor	3	28/49 (57,14% ± 6,79%)
	8	29/58 (50% ± 6,30%)
	9	27/43 (62,79% ± 7,08%)
	10	5/38 (13,16% ± 5,26%)
	11	23/55 (41,82% ± 6,39%)
	12	19/35 (54,29% ± 8,08%)
	13	39/61 (63,93% ± 5,90%)
	14	39/69 (56,52% ± 5,73%)
	15	16/56 (28,57% ± 5,80%)
	16	31/71 (43,66% ± 5,65%)

Tabla 39. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad de los *experimentos 7-12,* realizados con los datos de EMODB normalizados, para cada uno de los locutores, para 2 gausianas.

2 GAUS		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Locutor	3	Locutor	33/49 (67,35% ± 6,43%)	30/49 (61,22% ± 6,68%)	33/49 (67,35% ± 6,43%)
	8		28/58 (48,28% ± 6,30%)	25/58 (43,1% ± 6,24%)	27/58 (46,55% ± 6,29%)
	9		30/43 (69,77% ± 6,72%)	24/43 (55,81% ± 7,27%)	20/43 (46,51% ± 7,30%)
	10		14/38 (36,84% ± 7,51%)	14/38 (36,84% ± 7,51%)	17/38 (44,74% ± 7,74%)
	11		30/55 (54,55% ± 6,45%)	23/55 (41,82% ± 6,39%)	31/55 (56,36% ± 6,42%)
	12		18/35 (51,43% ± 8,11%)	18/35 (51,43% ± 8,11%)	16/35 (45,71% ± 8,08%)
	13		39/61 (63,93% ± 5,90%)	39/61 (63,93% ± 5,90%)	38/61 (62,3% ± 5,96%)
	14		43/69 (62,32% ± 5,60%)	46/69 (66,67% ± 5,45%)	52/69 (75,36% ± 4,98%)
	15		29/56 (51,79% ± 6,41%)	21/56 (37,5% ± 6,21%)	30/56 (53,57% ± 6,40%)
	16		43/71 (60,56% ± 5,57%)	33/71 (46,48% ± 5,68%)	54/71 (76,06% ± 4,86%)
N° Locutor	3	Neutra	30/49 (61,22% ± 6,68%)	29/49 (59,18% ± 6,74%)	30/49 (61,22% ± 6,68%)
	8		30/58 (51,72% ± 6,30%)	25/58 (43,1% ± 6,24%)	33/58 (56,9% ± 6,24%)
	9		31/43 (72,09% ± 6,57%)	24/43 (55,81% ± 7,27%)	27/43 (62,79% ± 7,08%)
	10		14/38 (36,84% ± 7,51%)	13/38 (34,21% ± 7,39%)	16/38 (42,11% ± 7,69%)
	11		28/55 (50,91% ± 6,47%)	20/55 (36,36% ± 6,23%)	28/55 (50,91% ± 6,47%)
	12		17/35 (48,57% ± 8,11%)	15/35 (42,86% ± 8,03%)	12/35 (34,29% ± 7,70%)
	13		39/61 (63,93% ± 5,90%)	37/61 (60,66% ± 6,00%)	35/61 (57,38% ± 6,08%)
	14		45/69 (65,22% ± 5,50%)	48/69 (69,57% ± 5,32%)	51/69 (73,91% ± 5,07%)
	15		30/56 (53,57% ± 6,40%)	22/56 (39,29% ± 6,27%)	35/56 (62,5% ± 6,21%)
	16		36/71 (50,7% ± 5,70%)	34/71 (47,89% ± 5,69%)	43/71 (60,56% ± 5,57%)

Tabla 40. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad del *experimento* 6, realizado con los datos de EMODB sin normalizar, para cada uno de los locutores, para 3 gausianas.

3 GAUS		Sin normalizar
N° Locutor	3	31/49 (63,27% ± 6,61%)
	8	18/58 (31,03% ± 5,83%)
	9	20/43 (46,51% ± 7,30%)
	10	7/38 (18,42% ± 6,04%)
	11	21/55 (38,18% ± 6,29%)
	12	12/35 (34,29% ± 7,70%)
	13	38/61 (62,3% ± 5,96%)
	14	47/69 (68,12% ± 5,39%)
	15	19/56 (33,93% ± 6,07%)
	16	30/71 (42,25% ± 5,63%)

Tabla 41. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad de los *experimentos 7-12*, realizados con los datos de EMODB normalizados, para cada uno de los locutores, para 3 gausianas.

3 GAUS		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Locutor	3	Locutor	32/49 (65,31% ± 6,53%)	31/49 (63,27% ± 6,61%)	32/49 (65,31% ± 6,53%)
	8		24/58 (41,38% ± 6,21%)	25/58 (43,1% ± 6,24%)	23/58 (39,66% ± 6,17%)
	9		22/43 (51,16% ± 7,32%)	26/43 (60,47% ± 7,16%)	25/43 (58,14% ± 7,22%)
	10		11/38 (28,95% ± 7,06%)	15/38 (39,47% ± 7,61%)	18/38 (47,37% ± 7,78%)
	11		30/55 (54,55% ± 6,45%)	22/55 (40% ± 6,34%)	31/55 (56,36% ± 6,42%)
	12		14/35 (40% ± 7,95%)	21/35 (60% ± 7,95%)	19/35 (54,29% ± 8,08%)
	13		40/61 (65,57% ± 5,84%)	33/61 (54,1% ± 6,13%)	43/61 (70,49% ± 5,61%)
	14		49/69 (71,01% ± 5,24%)	49/69 (71,01% ± 5,24%)	50/69 (72,46% ± 5,16%)
	15		33/56 (58,93% ± 6,31%)	29/56 (51,79% ± 6,41%)	33/56 (58,93% ± 6,31%)
	16		40/71 (56,34% ± 5,65%)	34/71 (47,89% ± 5,69%)	43/71 (60,56% ± 5,57%)
N° Locutor	3	Neutra	32/49 (65,31% ± 6,53%)	32/49 (65,31% ± 6,53%)	29/49 (59,18% ± 6,74%)
	8		30/58 (51,72% ± 6,30%)	21/58 (36,21% ± 6,06%)	33/58 (56,9% ± 6,24%)
	9		28/43 (65,12% ± 6,98%)	23/43 (53,49% ± 7,30%)	29/43 (67,44% ± 6,86%)
	10		13/38 (34,21% ± 7,39%)	17/38 (44,74% ± 7,74%)	17/38 (44,74% ± 7,74%)
	11		25/55 (45,45% ± 6,45%)	23/55 (41,82% ± 6,39%)	24/55 (43,64% ± 6,42%)
	12		14/35 (40% ± 7,95%)	17/35 (48,57% ± 8,11%)	12/35 (34,29% ± 7,7%)
	13		35/61 (57,38% ± 6,08%)	38/61 (62,3% ± 5,96%)	41/61 (67,21% ± 5,77%)
	14		47/69 (68,12% ± 5,39%)	47/69 (68,12% ± 5,39%)	50/69 (72,46% ± 5,16%)
	15		37/56 (66,07% ± 6,07%)	26/56 (46,43% ± 6,40%)	33/56 (58,93% ± 6,31%)
	16		46/71 (64,79% ± 5,44%)	35/71 (49,3% ± 5,70%)	45/71 (63,38% ± 5,49%)

Tabla 42. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad del *experimento 6,* realizado con los datos de EMODB sin normalizar, para cada uno de los locutores, para 4 gausianas.

4 GAUS		Sin normalizar
N° Locutor	3	31/49 (63,27% ± 6,61%)
	8	18/58 (31,03% ± 5,83%)
	9	20/43 (46,51% ± 7,30%)
	10	10/38 (26,32% ± 6,86%)
	11	24/55 (43,64% ± 6,42%)
	12	16/35 (45,71% ± 8,08%)
	13	35/61 (57,38% ± 6,08%)
	14	48/69 (69,57% ± 5,32%)
	15	20/56 (35,71% ± 6,15%)
	16	27/71 (38,03% ± 5,53%)

Tabla 43. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad de los *experimentos 7-12,* realizados con los datos de EMODB normalizados, para cada uno de los locutores, para 4 gausianas.

4 GAUS		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Locutor	3	Locutor	34/49 (69,39% ± 6,32%)	27/49 (55,1% ± 6,82%)	31/49 (63,27% ± 6,61%)
	8		21/58 (36,21% ± 6,06 %)	25/58 (43,1% ± 6,24%)	21/58 (36,21% ± 6,06%)
	9		24/43 (55,81% ± 7,27%)	26/43 (60,47% ± 7,16%)	24/43 (55,81% ± 7,27%)
	10		13/38 (34,21% ± 7,39%)	16/38 (42,11% ± 7,69%)	23/38 (60,53% ± 7,61%)
	11		30/55 (54,55% ± 6,45%)	24/55 (43,64% ± 6,42%)	33/55 (60% ± 6,34%)
	12		20/35 (57,14% ± 8,03%)	17/35 (48,57% ± 8,11%)	21/35 (60% ± 7,95%)
	13		39/61 (63,93% ± 5,90%)	35/61 (57,38% ± 6,08%)	44/61 (72,13% ± 5,51%)
	14		47/69 (68,12% ± 5,39%)	46/69 (66,67% ± 5,45%)	51/69 (73,91% ± 5,07%)
	15		34/56 (60,71% ± 6,27%)	30/56 (53,57% ± 6,40%)	32/56 (57,14% ± 6,35%)
	16		48/71 (67,61% ± 5,33%)	34/71 (47,89% ± 5,69%)	44/71 (61,97% ± 5,53%)
N° Locutor	3	Neutra	32/49 (65,31% ± 6,53%)	32/49 (65,31% ± 6,53%)	31/49 (63,27% ± 6,61%)
	8		28/58 (48,28% ± 6,30%)	20/58 (34,48% ± 5,99%)	34/58 (58,62% ± 6,21%)
	9		25/43 (58,14% ± 7,22%)	23/43 (53,49% ± 7,30%)	26/43 (60,47% ± 7,16%)
	10		17/38 (44,74% ± 7,74%)	16/38 (42,11% ± 7,69%)	16/38 (42,11% ± 7,69%)
	11		29/55 (52,73% ± 6,46%)	25/55 (45,45% ± 6,45%)	29/55 (52,73% ± 6,46%)
	12		14/35 (40% ± 7,95%)	17/35 (48,57% ± 8,11%)	13/35 (37,14% ± 7,84%)
	13		37/61 (60,66% ± 6%)	36/61 (59,02% ± 6,04%)	38/61 (62,3% ± 5,96%)
	14		46/69 (66,67% ± 5,45%)	51/69 (73,91% ± 5,07%)	49/69 (71,01% ± 5,24%)
	15		37/56 (66,07% ± 6,07%)	28/56 (50% ± 6,41%)	39/56 (69,64% ± 5,9%)
	16		43/71 (60,56% ± 5,57%)	37/71 (52,11% ± 5,69%)	46/71 (64,79% ± 5,44%)

Tabla 44. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad del *experimento 6,* realizado con los datos de EMODB sin normalizar, para cada uno de los locutores, para 5 gausianas.

5 GAUS		Sin normalizar
N° Locutor	3	29/49 (59,18% ± 6,74%)
	8	16/58 (27,59% ± 5,63%)
	9	20/43 (46,51% ± 7,30%)
	10	13/38 (34,21% ± 7,39%)
	11	22/55 (40% ± 6,34%)
	12	16/35 (45,71% ± 8,08%)
	13	34/61 (55,74% ± 6,11%)
	14	51/69 (73,91% ± 5,07%)
	15	23/56 (41,07% ± 6,31%)
	16	27/71 (38,03% ± 5,53%)

Tabla 45. Número de ficheros identificados correctamente, tasa de identificación y banda de fiabilidad de los *experimentos7-12,* realizados con los datos de EMODB normalizados, para cada uno de los locutores, para 5 gausianas.

5 GAUS		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Locutor	3	Locutor	24/49 (48,98% ± 6,86%)	31/49 (63,27% ± 6,61%)	29/49 (59,18% ± 6,74%)
	8		24/58 (41,38% ± 6,21%)	21/58 (36,21% ± 6,06%)	21/58 (36,21% ± 6,06%)
	9		24/43 (55,815 ± 7,27%)	24/43 (55,81% ± 7,27%)	21/43 (48,84% ± 7,32%)
	10		18/38 (47,37% ± 7,8%)	14/38 (36,84% ± 7,51%)	17/38 (44,74% ± 7,74%)
	11		34/55 (61,82% ± 6,29%)	27/55 (49,09% ± 6,47%)	30/55 (54,55% ± 6,45%)
	12		19/35 54,29% ± 8,08%)	16/35 (45,71% ± 8,08%)	22/35 (62,86% ± 7,84%)
	13		42/61 (68,85% ± 5,69%)	40/61 (65,57% ± 5,84%)	39/61 (63,93% ± 5,90%)
	14		50/69 (72,46% ± 5,16%)	42/69 (60,87% ± 5,64%)	52/69 (75,36% ± 4,98%)
	15		37/56 (66,07% ± 6,07%)	32/56 (57,14% ± 6,35%)	32/56 (57,14% ± 6,35%)
	16		41/71 (57,75% ± 5,63%)	32/71 (45,07% ± 5,67%)	47/71 (66,2% ± 5,39%)
N° Locutor	3	Neutra	29/49 (59,18% ± 6,74%)	31/49 (63,27% ± 6,61%)	36/49 (73,47% ± 6,05%)
	8		34/58 (58,62% ± 6,21%)	19/58 (32,76% ± 5,92%)	34/58 (58,62% ± 6,21%)
	9		25/43 (58,14% ± 7,22%)	27/43 (62,79% ± 7,08%)	29/43 (67,44% ± 6,86%)
	10		16/38 (42,11% ± 7,69%)	16/38 (42,11% ± 7,69%)	18/38 (47,37% ± 7,78%)
	11		31/55 (56,36% ± 6,42%)	25/55 (45,45% ± 6,45%)	27/55 (49,09% ± 6,47%)
	12		11/35 (31,43% ± 7,53%)	15/35 (42,86% ± 8,03%)	16/35 (45,71% ± 8,08%)
	13		37/61 (60,66% ± 6,00%)	44/61 (72,13% ± 5,51%)	36/61 (59,02% ± 6,04%)
	14		44/69 (63,77% ± 5,56%)	50/69 (72,46% ± 5,16%)	50/69 (72,46% ± 5,16%)
	15		35/56 (62,5% ± 6,21%)	27/56 (48,21% ± 6,41%)	37/56 (66,07% ± 6,07%)
	16		43/71 (60,56% ± 5,57%)	35/71 (49,3% ± 5,70%)	51/71 (71,83% ± 5,12%)

En las tablas anteriores (Tabla 36-*Tabla 45*) disponemos de muchos datos y resulta complicado extraer conclusiones a partir de ellas. Por ello, a continuación se muestran las tablas en las que se resume la media de los 10 locutores en función del número de gausianas utilizadas (Tabla 46 y Tabla 47).

El número total de ficheros en cada uno de los experimentos será el número total

de ficheros que tenemos en EMODB, **535**.

Tabla 46. Valor medio del número de ficheros identificados, la tasa de identificación y la banda de fiabilidad del *experimento 6,* realizado con los datos de EMODB sin normalizar, para los 10 locutores, en función del número de gausianas.

		Sin normalizar
Nº Gausianas	1	209 (43,55% ± 2,06%)
	2	225 (47,85% ± 2,07%)
	3	213 (45,42% ± 2,07%)
	4	222 (46,54% ± 2,07%)
	5	224 (46,92% ± 2,07%)

Tabla 47. Valor medio del número de ficheros identificados, la tasa de identificación y la banda de fiabilidad de los *experimentos 7-12,* realizados con los datos de EMODB normalizados, para los 10 locutores, en función del número de gausianas.

		Normalizado respecto a…	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
Nº Gausianas	1	Locutor	312 (58,32% ± 2,05%)	260 (48,6% ± 2,07%)	314 (58,69% ± 2,04%)
	2		307 (57,38% ± 2,05%)	273 (51,03% ± 2,07%)	318 (59,44% ± 2,04%)
	3		295 (55,14% ± 2,06%)	285 (53,27% ± 2,07%)	317 (59,25% ± 2,04%)
	4		310 (57,94% ± 2,05%)	280 (52,34% ± 2,07%)	324 (60,56% ± 2,03%)
	5		313 (58,5% ± 2,04%)	279 (52,15% ± 2,07%)	310 (57,94% ± 2,05%)
	1	Neutra	308 (57,57% ± 2,05%)	261 (48,78% ± 2,07%)	314 (58,69% ± 2,04%)
	2		300 (56,07% ± 2,06%)	267 (49,91% ± 2,08%)	310 (57,94% ± 2,05%)
	3		307 (57,38% ± 2,05%)	279 (52,15% ± 2,07%)	313 (58,5% ± 2,05%)
	4		308 (57,57% ± 2,05%)	285 (53,27% ± 2,07%)	321 (60% ± 2,03%)
	5		305 (57,01% ± 2,05%)	289 (54,02% ± 2,07%)	334 (62,43% ± 2,01%)

Analizamos la mayor mejora obtenida al aplicar las técnicas de normalización, en función del número de gausianas empleado en el modelo:

- *1 gausiana:* $\mathbf{\Delta E_{acc}}$ = (58,69% - 43,55%) = **15,14%**
- *2 gausianas:* $\mathbf{\Delta E_{acc}}$ = (59,44% - 47,85%) = **11,59%**
- *3 gausianas:* $\mathbf{\Delta E_{acc}}$ = (59,25% - 45,42%) = **13,33%**

- *4 gausianas:* $\mathbf{\Delta E_{acc}}$ = (60,56% - 46,54%) = **14,02%**
- *5 gausianas:* $\mathbf{\Delta E_{acc}}$ = (62,43% - 46,92%) = **15,51%**

Observamos que en todos los casos obtenemos una mejora mayor del 10% al normalizar. La normalización con la que mejores resultados obtenemos es, en general, en la que normalizamos respecto a la media y la varianza de la voz del locutor. Pero hay excepciones, para el caso de 1 gausiana se obtienen los mismos resultados para esa normalización y para el caso en el que normalicemos respecto a la voz neutra del locutor. Y para el caso de 5 gausianas, los resultados son mejores cuando normalizamos con la media y la varianza respecto a la voz neutra del locutor.

Fijándonos en las bandas de fiabilidad de los resultados obtenidos en el caso en el que utilizamos vectores normalizados, concluimos que:

- Si normalizamos los vectores respecto a la voz del locutor o respecto a la voz neutra del locutor, los resultados obtenidos son similares. Debemos tener en cuenta que la media y la varianza estimada sobre la voz neutra se basa en un menor número de datos.
- Los resultados obtenidos tanto si empleamos la media, como si empleamos la media y la varianza para normalizar, se mueven en valores similares, siendo ligeramente superiores cuando utilizamos la media y la varianza.
- La mejora sí que es estadísticamente fiable cuando utilizamos para la normalización solamente la media, o la media y la varianza, en comparación a cuando solamente utilizamos la varianza.

6.2.3.Análisis de la tasa de identificación para cada emoción

A continuación compararemos las tasas de identificación para cada emoción cuando utilizamos 1 gausiana para los casos en los que realizamos los experimentos con vectores de características sin normalizar y cuando realizamos los experimentos con vectores normalizados, empleando las normalizaciones para las que hemos visto en el apartado anterior que obteníamos mejores resultados, es decir, normalizando con media y varianza tanto respecto a la voz del locutor como a la voz neutra de dicho locutor.

Las siguientes tablas (Tabla 48-*Tabla 50*) muestran las tasas de identificación para

cada emoción, para cada uno de estos experimentos.

Tabla 48. Tasa de identificación para cada emoción del *experimento 6,* realizado sobre EMODB con vectores sin normalizar.

	EMOCIÓN IDENTIFICADA						
EMOCIÓN INTERPRETADA	Alegría	Enfado	Aburrimiento	Tristeza	Neutro	Asco	Miedo
Alegría	32,39% (± 5,33%)	23,94%	1,41%		2,82%	22,54%	16,90%
Enfado	14,96%	62,20% (± 4,13%)	0,79%		0,79%	1,57%	19,69%
Aburrimiento	1,23%	2,47%	27,16% (± 4,74%)	19,75%	19,75%	22,22%	7,41%
Tristeza			8,06%	83,87% (± 4,48%)	1,61%	1,61%	4,84%
Neutro			31,65%	17,72%	27,85% (± 4,84%)	15,19%	7,59%
Asco	8,70%	13,04%	13,04%	4,35%	13,04%	39,13% (± 6,91%)	8,70%
Miedo	1,45%	15,94%	2,90%	14,49%	20,29%	20,29%	24,64% (± 4,98%)
PRECISIÓN	55,15%	52,90%	31,95%	59,83%	32,33%	31,93%	27,45%

Tabla 49. Tasa de identificación para cada emoción del *experimento 9,* en el que empleamos características de EMODB normalizadas con media y varianza estimadas respecto al locutor.

	EMOCIÓN IDENTIFICADA						
EMOCIÓN INTERPRETADA	Alegría	Enfado	Aburrimiento	Tristeza	Neutro	Asco	Miedo
Alegría	42,25% (± 5,63%)	26,76%			1,41%	5,63%	23,94%
Enfado	13,39%	71,65% (± 3,84%)				0,79%	14,17%
Aburrimiento			43,21% (± 5,28%)	12,35%	25,93%	12,35%	6,17%
Tristeza			1,61%	95,16% (± 2,62%)	3,23%		
Neutro			36,71%	3,80%	43,04% (± 5,35%)	11,39%	5,06%
Asco	2,17%		2,17%		8,70%	73,91% (± 6,22%)	13,04%
Miedo	8,70%	13,04%	4,35%	7,25%	10,14%	11,59%	44,93% (± 5,75%)
PRECISIÓN	63,52%	64,29%	49,07%	80,26%	46,55%	63,90%	41,87%

Tabla 50. Tasa de identificación para cada emoción del *experimento 12,* en el que empleamos características de EMODB normalizadas con media y varianza estimada respecto a la neutra del locutor.

	EMOCIÓN IDENTIFICADA						
EMOCIÓN INTERPRETADA	Alegría	Enfado	Aburrimiento	Tristeza	Neutro	Asco	Miedo
Alegría	36,62% (± 5,49%)	29,58%		1,41%	5,63%	5,63%	21,13%
Enfado	15,75%	65,35% (± 4,05%)					18,90%
Aburrimiento			40,74% (± 5,24%)	1,23%	50,62%	6,17%	1,23%
Tristeza			1,61%	85,48% (± 4,3%)	12,90%		
Neutro			20,25%		79,75% (± 4,34%)		
Asco	2,17%	2,17%	10,87%	2,17%	8,70%	63,04% (± 6,83%)	10,87%
Miedo	7,25%	17,39%	14,49%	4,35%	7,25%	10,14%	39,13% (± 5,64%)
PRECISIÓN	59,27%	57,08%	46,32%	90,32%	48,38%	74,18%	42,88%

En los tres casos estudiados se cumple que la emoción que mejor se identifica es la tristeza, obteniendo la mayor tasa de identificación (95,16%) para el caso en el que utilizamos vectores de características normalizados con la media y la varianza estimada respecto a la voz del locutor.

La segunda emoción que mejor se identifica es distinta en cada uno de los casos: el enfado (62,2%), para el caso en el que utilizamos vectores sin normalizar; el asco (73,91%), cuando normalizamos los vectores respecto a la voz del locutor; y la voz neutra (79,75%), cuando normalizamos respecto a la voz neutra del locutor.

La emoción que peor se identifica también depende de cada caso. Así, cuando no normalizados, el miedo es el que peor se identifica (24,64%); cuando normalizamos tanto respecto a la voz del locutor como respecto a la voz neutra de dicho locutor, la alegría es la emoción que peor se identifica (con un 42,25% y un 36,62%, respectivamente).

Analizamos a continuación el comportamiento de la voz neutra en estos experimentos:

- En el caso en el que utilizamos vectores de características sin normalizar, se confunde con el aburrimiento (31,65%).
- Al normalizar los vectores respecto al locutor se identifica en un 43,04% de los casos, pero en un 36,71% de los casos se sigue confundiendo con el aburrimiento.
- Al normalizar respecto a la neutra del locutor conseguimos identificarla con una tasa elevada (79,75%). Su precisión no aumenta mucho debido a que el aburrimiento se confunde mucho con la neutra (50,62%).

Analizamos la mejora obtenida para cada emoción empleando vectores normalizados, así como la correspondiente banda de fiabilidad, para comprobar si la mejora es o no significativa; y la mejora relativa del error:

- Respecto a la voz del locutor:

  - *Alegría*: ΔE_{acc} = (42,25% - 32,39%) = **9,86%**

    Banda de fiabilidad = 5,33% + 5,63% = **10,96%**

    Mejora no significativa

    Mejora relativa del error = **14,58%**

  - *Enfado*: ΔE_{acc} = (71,65% - 62,2%) = **9,45%**

    Banda de fiabilidad = 4,13% + 3,84% = **7,97%**

    Mejora relativa del error = **25%**

  - *Aburrimiento*: ΔE_{acc} = (43,21% - 27,16%) = **16,05%**

    Banda de fiabilidad = 4,74% + 5,28% = **10,02%**

    Mejora relativa del error = **22,03%**

  - *Tristeza*: ΔE_{acc} = (95,16% - 83,87%) = **11,29%**

    Banda de fiabilidad = 4,48% + 2,62% = **7,1%**

    Mejora relativa del error = **70%**

  - *Neutro*: ΔE_{acc} = (43,04% - 27,85%) = **15,19%**

    Banda de fiabilidad = 4,84% + 5,35% = **10,19%**

    Mejora relativa del error = **21,05%**

  - *Asco*: ΔE_{acc} = (73,91% - 39,13%) = **34,78%**

    Banda de fiabilidad = 6,91% + 6,22% = **13,13%**

    Mejora relativa del error = **57,14%**

  - *Miedo*: ΔE_{acc} = (44,93% - 24,64%) = **20,29%**

    Banda de fiabilidad = 4,98% + 5,75% = **10,73%**

    Mejora relativa del error = **26,92%**

En media se obtiene entre un 10% y un 15% de mejora absoluta al normalizar, aunque como observamos, algunas emociones sufren una mejora mucho mayor, como el asco y el miedo. Excepto en el caso de la alegría, las mejoras obtenidas son significativas. La mayor mejora relativa del error la obtiene la tristeza.

- Respecto a la voz neutra del locutor:

  - *Alegría*: $\mathbf{\Delta E_{acc}}$ = (36,62% - 32,39%) = **4,23%**

    Banda de fiabilidad = 5,33% + 5,49% = **10,82%**

    Mejora relativa del error = **6,26%**

    } Mejora no significativa

  - *Enfado*: $\mathbf{\Delta E_{acc}}$ = (65,35% - 62,20%) = **3,15%**

    Banda de fiabilidad = 4,13% + 4,05% = **8,18%**

    Mejora relativa del error = **12%**

    } Mejora no significativa

  - *Aburrimiento*: $\mathbf{\Delta E_{acc}}$ = (40,74% - 27,16%) = **13,58%**

    Banda de fiabilidad = 4,74% + 5,24% = **9,98%**

    Mejora relativa del error = **18,64%**

  - *Tristeza*: $\mathbf{\Delta E_{acc}}$ = (85,48% - 83,87%) = **1,61%**

    Banda de fiabilidad = 4,48% + 4,3% = **8,78%**

    Mejora relativa del error = **9,98%**

    } Mejora no significativa

  - *Neutro*: $\mathbf{\Delta E_{acc}}$ = (79,75% - 27,85%) = **51,90%**

    Banda de fiabilidad = 4,84% + 4,34% = **9,18%**

    Mejora relativa del error = **71,93%**

  - *Asco*: $\mathbf{\Delta E_{acc}}$ = (63,04% - 39,13%) = **23,91%**

    Banda de fiabilidad = 6,91% + 6,83% = **13,74%**

    Mejora relativa del error = **39,28%**

- *Miedo*: $\mathbf{\Delta E_{acc}}$ = (39,13% - 24,64%) = **14,49%**

  Banda de fiabilidad = 4,98% + 5,64% = **10,96%**

  Mejora relativa del error = **19,23%**

En este caso también obtenemos una mejora respecto al caso en el que empleamos vectores sin normalizar, pero ésta es menor que la obtenida normalizando respecto a toda la voz del locutor, sobre todo para la alegría, el enfado y la tristeza, que obtenemos mejoras no significativas. Esto se puede deber a que la media y la varianza estimada respecto a la voz neutra, se basan en menos datos. La neutra es la emoción que consigue la mayor mejora absoluta y la mayor mejora relativa del error, por lo que podemos afirmar que se trata de la emoción a la que mejor le sienta la normalización respecto a la voz neutra del locutor.

En los tres experimentos se cumple que:

- El enfado nunca se confunde con la tristeza.
- La tristeza nunca se confunde con la alegría ni con el enfado.
- La neutra nunca se confunde con la alegría ni con el enfado.
- La alegría se confunde muy poco con el aburrimiento y la tristeza (sólo en el caso en el que no normalizamos se confunde en un 1,41% con el aburrimiento, y en el caso en el que normalizamos respecto a la voz neutra, un 1,41% con la tristeza).

Realizando los experimentos con vectores normalizados tanto respecto a la voz del locutor como a su voz neutra conseguimos aumentar el número de emociones que no se confunden con otras. Así, en esos casos, el aburrimiento nunca se confunde con la alegría ni con el enfado, y la tristeza nunca se confunde ni con el asco ni con el miedo. Esto lo podemos ver claramente observando la precisión de cada emoción para cada uno de los experimentos. Aunque ya aparecía la precisión de cada emoción en las tablas anteriores (Tabla 48-*Tabla 50*), en la siguiente tabla (Tabla 51) se muestra la precisión de cada una de las emociones para los tres tipos de experimentos que estamos analizando.

Tabla 51. Precisión de cada emoción para los experimentos realizados con las características de EMODB sin normalizar y normalizando respecto a la voz del locutor y respecto a la voz neutra.

	Alegría	Enfado	Aburrimiento	Tristeza	Neutro	Asco	Miedo
Sin normalizar	55,15%	52,90%	31,95%	59,83%	32,33%	31,93%	27,45%
Media + Varianza (locutor)	63,52%	64,29%	49,07%	80,26%	46,55%	63,90%	41,87%
Media + Varianza (neutra)	59,27%	57,08%	46,32%	90,32%	48,38%	74,18%	42,88%

Podemos observar como la precisión de todas las emociones aumenta cuando normalizamos (ya sea respecto al locutor o respecto a la neutra). Para la alegría, el enfado y el aburrimiento, se obtiene mayor precisión normalizando los vectores respecto a la voz del locutor. Para la tristeza, la neutra, el asco y el miedo, se obtiene mayor precisión normalizando respecto a la voz neutra del locutor.

Para el asco es para la emoción que más aumenta la precisión al normalizar, mejorando de un 31,93% a un 74,18% (aumenta un 42,25%). En la tristeza también aumenta considerablemente, mejorando de un 59,83% a un 90,32% (aumenta un 30,49%).

6.2.4.Conclusiones de los experimentos de identificación de emociones sobre EMODB

Una vez analizados los resultados obtenidos en los distintos experimentos de identificación realizados con la base de datos EMODB con características basadas en MFCC, las conclusiones a las que llegamos son:

- Al normalizar obtenemos mejoras importantes, mayores del 10%, en todos los casos. En general, la mejora aumenta al aumentar el número de gausianas utilizado, pero los resultados obtenidos con 1 gausiana son también muy buenos, similares a los obtenidos con 5 gausianas (15,14% frente a 15,51%).

- La normalización con la que obtenemos mejores resultados es aquella en la que normalizamos respecto a la media y la varianza de la voz del locutor. Como en EMODB disponemos de ficheros grabados por diez locutores diferentes,

normalizando respecto a la voz del locutor, conseguimos eliminar las características propias y mantener únicamente las características comunes a todos los locutores para interpretar una cierta emoción.

- La emoción que mejor se identifica, en los tres casos analizados (con vectores sin normalizar y normalizándolos respecto a la voz del locutor y respecto a su voz neutra) es la tristeza, obteniéndose la mayor tasa de identificación (95,16%) cuando utilizamos vectores normalizados con media y varianza estimada a partir de todos los datos de la voz del locutor. El enfado y el asco también obtienen tasas de identificación buenas (71,65% y 73,91%, respectivamente, normalizando respecto a la voz del locutor).

- Las emociones que peor se identifican son el miedo y la alegría.

- La voz neutra se confunde con el aburrimiento, pero conseguimos que se identifique, con un 79,75%, cuando normalizamos respecto a la voz neutra del locutor. Es decir, la mejora obtenida al normalizar es elevada (51,9%).

- En el caso del asco, también obtenemos una elevada mejora absoluta al normalizar (34,78%), mientras que la mayor mejora del error relativo al normalizar es para la tristeza (70%).

- La precisión de todas las emociones aumenta cuando normalizamos. Para las emociones para las que más aumenta es para el asco y la tristeza, normalizando respecto a la neutra del locutor.

- En todos los casos se cumple que:
    - El enfado nunca se confunde con la tristeza.
    - La tristeza y la neutra nunca se confunden con la alegría, ni con el enfado.
    - La alegría nunca se confunde con el aburrimiento, ni con la tristeza.

- Al normalizar, conseguimos que el aburrimiento nunca se confunda con la alegría, ni con el enfado; y que la tristeza nunca se confunda con el asco, ni con el miedo.

6.3.Identificación de emociones entre distintos idiomas

En este apartado describiremos los experimentos en los cuales entrenamos con datos de una de las bases de datos (SES y EMODB) y clasificamos datos de la otra. Es decir, en los experimentos realizados en este apartado obtendremos modelos basados en un idioma, y clasificaremos ficheros de otro idioma distinto.

La razón por la que hemos realizado este tipo de identificación en la que usamos distintos idiomas es para intentar descubrir posibles similitudes o diferencias entre la interpretación de las emociones en castellano y en alemán, que son los idiomas de las dos bases de datos disponibles en el grupo.

Al igual que en los experimentos realizados sólo con SES o EMODB, en este tipo de experimentos podemos aplicar o no técnicas de normalización, respecto a la voz del locutor o sólo respecto a la voz neutra de dicho locutor, con la media y/o la varianza.

Según qué base de datos empleemos en el entrenamiento y qué base de datos empleemos en la clasificación, y si utilizamos todas las emociones o sólo las comunes a ambas bases de datos, obtenemos tres tipos diferentes de experimentos, que describimos en los siguientes apartados.

6.3.1.Entrenamiento con datos de SES y clasificación de datos de EMODB, utilizando todas las emociones

Los experimentos realizados en este apartado consistirán en obtener las tasas de identificación de los datos disponibles en lengua alemana, a partir de unos modelos basados en lengua castellana. Es decir, entrenaremos nuestro modelo con vectores de características obtenidos a partir de los datos de SES y clasificaremos los vectores de características obtenidos a partir de los datos de EMODB.

1.1.1.25.Descripción de los experimentos.

En la base de datos de SES tenemos 4 emociones (alegría, enfado, sorpresa y tristeza) más la neutra, y en la de EMODB tenemos 6 emociones (alegría, enfado, aburrimiento, tristeza, asco y miedo) más la neutra. Por lo tanto, en el entrenamiento realizado con datos de SES obtendremos el modelo para la alegría, el enfado, la sorpresa, la tristeza y el estado neutro. A continuación pasaremos a clasificar todos los ficheros de las distintas emociones de EMODB, y podremos analizar con que emociones de SES se identifican las tres emociones que no tienen correspondencia (aburrimiento, asco y miedo).

Los posibles experimentos que obtenemos según normalicemos o no los vectores de entrenamiento y clasificación, y en función de la normalización aplicada en su caso, son los siguientes:

- EXPERIMENTO 13: *sin normalizar los vectores de características.*
- EXPERIMENTO 14: *CMN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 15: *CVN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 16: *CMN + CVN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 17: *CMN (normalizando respecto a la voz neutra del locutor).*
- EXPERIMENTO 18: *CVN (normalizando respecto a la voz neutra del locutor).*
- EXPERIMENTO 19: *CMN + CVN (normalizando respecto a la voz neutra del locutor).*

1.1.1.26.Resultados de los experimentos

En las siguientes tablas (Tabla 52 y Tabla 53) se muestran el número de ficheros identificados, así como la tasa de identificación y su correspondiente banda de fiabilidad, para cada uno de los experimentos en los que entrenamos con datos de SES y clasificamos los ficheros de EMODB, tanto de las emociones comunes como de las que no son comunes a ambas bases de datos. Se muestran en diferentes tablas los resultados obtenidos para el experimento en el que utilizamos vectores sin normalizar (*experimento 13)* y los resultados para los experimentos con vectores normalizados

(*experimentos 14-19)*.

El número total de ficheros de clasificación que tenemos en todos los experimentos de este apartado es el número total de ficheros disponibles en EMODB, es decir, **535**.

Tabla 52. Número de ficheros identificados, tasa de identificación y banda de fiabilidad del *experimento 13,* en el que entrenamos con SES y clasificamos EMODB, sin normalizar, con todas las emociones, en función del número de gausianas.

		Sin normalizar
N° Gausianas	1	121 (22,62% ± 1,74%)
	2	113 (21,12% ± 1,69%)
	3	126 (23,55% ± 1,76%)
	4	115 (21,5% ± 1,71%)
	5	115 (21,5% ± 1,71%)

Tabla 53. Número de ficheros identificados, tasa de identificación y banda de fiabilidad de los *experimentos 14-19,* en los que entrenamos con SES y clasificamos EMODB, normalizando, con todas las emociones, en función del número de gausianas.

		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Gausianas	1	Locutor	130 (24,3% ± 1,78%)	114 (21,31% ± 1,70%)	126 (23,55% ± 1,76%)
	2		109 (20,37% ± 1,67%)	134 (25,05% ± 1,80%)	152 (28,41% ± 1,87%)
	3		123 (22,99% ± 1,75%)	130 (24,3% ± 1,78%)	138 (25,79% ± 1,82%)
	4		100 (18,69% ± 1,62%)	130 (24,3% ± 1,78%)	134 (25,05% ± 1,80%)
	5		104 (19,44% ± 1,64%)	136 (25,42% ± 1,81%)	123 (22,99% ± 1,75%)
N° Gausianas	1	Neutra	137 (25,61% ± 1,81%)	129 (24,11% ± 1,78%)	146 (27,29% ± 1,85%)
	2		79 (14,77% ± 1,47%)	132 (24,67% ± 1,79%)	152 (28,41% ± 1,87%)
	3		104 (19,44% ± 1,64%)	128 (23,93% ± 1,77%)	133 (24,86% ± 1,79%)
	4		81 (15,14% ± 1,49%)	132 (24,67% ± 1,79%)	143 (26,73% ± 1,84%)
	5		77 (14,39% ± 1,46%)	141 (26,36% ± 1,83%)	124 (23,18% ± 1,75%)

Las tasas de identificación obtenidas para estos experimentos son muy bajas, ya que hemos clasificado todas las emociones de EMODB, mientras que sólo son cuatro de ellas las que coinciden con SES, que son de las que hemos obtenido un modelo en el entrenamiento. Más adelante analizaremos la tasa de identificación para cada una de las emociones y podremos comprobar con que emociones se identifican aquellas de las que no hemos obtenido un modelo al entrenar.

En estos experimentos tenemos 277.513 vectores de entrenamiento frente a los 146.047 vectores de clasificación, es decir, tenemos una proporción aproximada de un 65% en entrenamiento y un 35% en clasificación, que es adecuado.

La mayor tasa de identificación se obtiene cuando normalizamos con media y varianza respecto a la voz neutra del locutor, salvo para el caso de utilizar 3 gausianas, que se obtiene cuando normalizamos con media y varianza, pero respecto a toda la voz del locutor, y en el caso de utilizar 5 gausianas, que se obtiene cuando normalizamos sólo con la varianza respecto a la voz neutra.

La normalización que peor funciona es aquella en la que estimamos la media ya sea a partir de toda la voz del locutor o sólo a partir de su voz neutra. Al entrenar y clasificar con los vectores con estos tipos de normalización, se obtienen tasas de identificación inferiores a las que obtenemos cuando utilizamos los vectores sin normalizar, excepto para 1 gausiana.

Las diferencias entre las tasas de identificación obtenidas cuando empleamos los vectores normalizados respecto a la varianza o respecto a la media y la varianza no son estadísticamente fiables, ya que se solapan sus bandas de fiabilidad, por lo que no podemos concluir que una normalización sea mejor que la otra.

A continuación, analizamos la mejora obtenida al utilizar vectores normalizados en función del número de gausianas empleado:

- *1 gausiana*: $\mathbf{\Delta E_{acc}}$ = (27,29% - 22,62%) = **4,67%**
- *2 gausianas*: $\mathbf{\Delta E_{acc}}$ = (28,41% - 21,12%) = **7,29%**
- *3 gausianas*: $\mathbf{\Delta E_{acc}}$ = (25,79% - 23,55%) = **2,24%**

- *4 gausianas*: $\mathbf{\Delta E_{acc}}$ = (26,73% - 21,5%) = **5,23%**
- *5 gausianas*: $\mathbf{\Delta E_{acc}}$ = (26,36% - 21,5%) = **4,86%**

Podemos observar que en ninguno de los casos se obtiene una importante mejora al normalizar. De hecho, cuando utilizamos 3 gausianas, la posible mejora que obtenemos (2,24%) no es estadísticamente fiable. La mayor mejora se obtiene para el caso en el que utilizamos 2 gausianas, que también es el caso en el que conseguimos la mayor tasa de identificación (28,41%).

No podemos establecer cual es el número óptimo de gausianas que debemos utilizar en nuestro sistema, ya que dependiendo de la normalización que apliquemos a las características de nuestro sistema, será mejor un número de gausianas u otro.

- Cuando utilizamos los vectores normalizados respecto a la media, la mayor tasa de identificación se obtiene cuando utilizamos 1 gausiana.
- Utilizando vectores normalizados respecto a la varianza, la mayor tasa de identificación se obtiene cuando utilizamos 5 gausianas.
- Utilizando vectores normalizados respecto a la media y la varianza, la tasa de identificación aumenta al pasar de utilizar 1 gausiana a 2, pero a partir de ahí, disminuye.

1.1.1.27. Análisis de la tasa de identificación para cada emoción

En este apartado vamos a analizar las tasas de identificación para cada emoción para los experimentos en los que entrenamos con datos SES y clasificamos datos de EMODB, considerando todas las emociones de ambas bases de datos, en el caso en el que utilizamos 1 gausiana. Mostramos en distintas tablas los resultados obtenidos cuando utilizamos vectores sin normalizar (Tabla 54) y cuando utilizamos vectores normalizados respecto a la media y la varianza estimada a partir de la voz del locutor (Tabla 55) y estimados a partir de la voz neutra de dicho locutor (Tabla 56), ya que en el apartado anterior hemos visto que para este tipo de normalización es para la que obteníamos mejores resultados (aunque estos resultados eran similares a los obtenidos al emplear vectores normalizados sólo respecto a la varianza, de hecho, sus bandas de fiabilidad se solapaban).

En las siguientes tablas (Tabla 54-*Tabla 56)* se muestra, además de la tasa de identificación para cada emoción, la precisión de cada una de las emociones.

Tabla 54. Tasas de identificación para cada emoción del *experimento 13,* en el que entrenamos con SES y clasificamos con EMODB con vectores sin normalizar, con todas las emociones.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	9,86% (± 3,4%)		90,14%		
Enfado	2,36%	1,57% (± 1,06%)	96,06%		
Aburrimiento	4,94%	3,70%	53,09%	20,99%	17,28%
Tristeza	3,23%		8,06%	85,48% (±4,29%)	3,23%
Neutro	18,99%	3,80%	43,04%	13,92%	20,25% (± 4,34%)
Asco	10,87%	2,17%	78,26%	6,52%	2,17%
Miedo	14,49%	7,25%	66,67%	11,59%	
PRECISIÓN	15,23%	8,51%		61,72%	47,17%

Tabla 55. Tasas de identificación para cada emoción del *experimento 16,* en el que entrenamos con SES y clasificamos con EMODB con vectores normalizados con media y varianza respecto al locutor, con todas las emociones.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	32,39% (± 5,33%)	15,49%	50,70%		1,41%
Enfado	11,02%	7,87% (± 2,29%)	81,10%		
Aburrimiento	4,94%	9,88%	3,70%	44,44%	37,04%
Tristeza				100%	
Neutro	11,39%	10,13%	1,27%	41,77%	35,44% (5,17%)
Asco	45,65%	19,57%	6,52%	6,52%	21,74%
Miedo	13,04%	20,29%	46,38%	18,84%	1,45%
PRECISIÓN	27,35%	9,46%		47,26%	36,51%

Tabla 56. Tasas de identificación para cada emoción del *experimento* 19 en el que entrenamos con SES y clasificamos con EMODB con vectores normalizados con media y varianza respecto a la voz neutra del locutor, con todas las emociones.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	26,76% (± 5,04%)	5,63%	49,30%		18,31%
Enfado	16,54%	6,30% (± 2,07%)	74,80%		2,36%
Aburrimiento	1,23%	1,23%	1,23%	3,70%	92,59%
Tristeza				64,52% (± 5,83%)	35,48%
Neutro		1,27%			98,73% (± 1,21%)
Asco	23,91%		21,74%		54,35%
Miedo	2,90%	10,14%	53,62%		33,33%
PRECISIÓN	37,51%	25,63%		94,57%	29,46%

Las conclusiones que obtenemos al observar los resultados mostrados en estas tablas son las siguientes:

- La emoción que mejor se identifica es la tristeza, identificándose siempre en el caso de normalizar los vectores respecto a la media y la varianza estimadas a partir de la voz del locutor.
- Al utilizar vectores normalizados respecto a la voz neutra del locutor, conseguimos aumentar considerablemente la tasa de identificación de la neutra, de un 25,25% a un 98,73%. En el caso en el que utilizamos vectores sin normalizar, la neutra se confunde con la sorpresa (43,04%).
- Todas las emociones, salvo la tristeza, se confunden con la sorpresa, cuando utilizamos vectores sin normalizar. Esto lo podemos asociar a que las emociones en alemán se parecen a la interpretación que hace de la sorpresa el actor de SES. Al normalizar se reduce la confusión de todas las emociones con la sorpresa y algunas de ellas comienzan a confundirse con otras emociones:
    - El aburrimiento se confunde con la tristeza (44,44% normalizando respecto a la voz del locutor) y con la neutra (37,04% normalizando respecto a la voz del locutor y 92,59% normalizando respecto a su voz

neutra).

  - El asco se confunde con la alegría al normalizar respecto a la voz del locutor (45,65%) y con la neutra al normalizar respecto a su voz neutra (54,35%).
  - La neutra se confunde con la tristeza al normalizar respecto a la voz del locutor (41,77%).

- Conseguimos que las emociones se confundan menos unas con otras cuando empleamos vectores normalizados respecto a la voz neutra del locutor, que cuando empleamos los normalizados respecto a la voz del locutor o los vectores sin normalizar. En todos los casos se cumple que la alegría y el enfado nunca se confunden con la tristeza, y la tristeza nunca se confunde con el enfado.
- Como consecuencia del punto anterior, en principio se debe cumplir que la precisión de las distintas emociones cuando utilizamos vectores normalizados respecto a la voz neutra del locutor sea mayor que cuando utilizamos los normalizados respecto a la voz del locutor o los vectores sin normalizar. Observando la tabla, vemos que esto se cumple salvo para el caso de la neutra, cuya precisión es mayor al emplear vectores sin normalizar (47,17% frente a 36,51% y 29,46%). Esto se debe a que al normalizar, el aburrimiento se confunde mucho con la neutra, las tasas de confusión son 37,04% y 92,59%.
- La tristeza tiene la precisión más elevada (94,57%) cuando utilizamos vectores normalizados respecto a la voz neutra.

La sorpresa no se corresponde con ninguna emoción de la base de datos de EMODB, por lo que no hemos podido calcular su precisión.

6.3.2. Entrenamiento con datos de SES y clasificación de datos de EMODB, utilizando sólo las emociones comunes

Una vez realizados los experimentos entrenando con datos de SES y clasificando datos de EMODB, con todas las emociones disponibles en ambas bases de datos, en los que hemos analizado con qué emociones se confunden las emociones que no tienen correspondencia y, por lo tanto, de las que no hemos extraído un modelo en el entrenamiento; abordaremos en este apartado la realización de los experimentos en los que sólo consideremos las emociones comunes a ambas bases de datos, para evaluar el comportamiento de nuestro sistema al clasificar emociones con modelos extraídos a partir de emociones de un idioma diferente.

1.1.1.28. Descripción de los experimentos.

Los experimentos que realizaremos en este apartado son similares a los del apartado anterior (*apartado 1.1.1.24*), pero utilizando sólo las emociones comunes en ambas bases de datos, es decir, la alegría, el enfado, la tristeza y la voz neutra. Por lo tanto, en función de si aplicamos o no normalización a los vectores de características empleados, así como en función del tipo de normalización empleado, los posibles experimentos que obtenemos son los siguientes:

- EXPERIMENTO 20: *sin normalizar los vectores de características.*
- EXPERIMENTO 21: *CMN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 22: *CVN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 23: *CMN + CVN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 24: *CMN (normalizando respecto a la voz neutra del locutor).*
- EXPERIMENTO 25: *CVN (normalizando respecto a la voz neutra del locutor).*
- EXPERIMENTO 26: *CMN + CVN (normalizando respecto a la voz neutra del locutor).*

1.1.1.29. Resultado de los experimentos

En las siguientes tablas (Tabla 57 y Tabla 58) se muestran el número de ficheros identificados, así como la tasa de identificación y su correspondiente banda de fiabilidad, para cada uno de los experimentos de identificación en los que entrenamos con datos de las emociones comunes de SES y clasificamos los ficheros de las emociones comunes de EMODB. Se muestran en diferentes tablas los resultados obtenidos para el experimento en el que utilizamos vectores sin normalizar (*experimento 20)* y los resultados para los experimentos con vectores normalizados (*experimentos 21 -26)*.

El número total de ficheros de clasificación que tenemos en todos los experimentos de este apartado es **339**.

Tabla 57. Número de ficheros identificados, tasa de identificación y banda de fiabilidad del *experimento* 20, en el que entrenamos con datos de SES y clasificamos datos de EMODB, sin normalizar, sólo con las emociones comunes, en función del número de gausianas.

		Sin normalizar
N° Gausianas	1	143 (42,18% ± 2,57%)
	2	144 (42,48% ± 2,58%)
	3	166 (48,97% ± 2,61%)
	4	167 (49,26% ± 2,61%)
	5	170 (50,15% ± 2,61%)

Tabla 58. Número de ficheros identificados, tasa de identificación y banda de fiabilidad de los *experimentos 21-26,* en los que entrenamos con datos de SES y clasificamos datos de EMODB, normalizando, sólo con las emociones comunes, en función del número de gausianas.

		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Gausianas	1	Locutor	163 (48,08% ± 2,61%)	159 (46,9% ± 2,60%)	174 (51,33% ± 2,61%)
	2		156 (46,02% ± 2,60%)	166 (48,97% ± 2,61%)	197 (58,11% ± 2,57%)
	3		174 (51,33% ± 2,61%)	172 (50,74% ± 2,61%)	181 (53,39% ± 2,60%)
	4		166 (48,97% ± 2,61%)	167 (49,26% ± 2,61%)	170 (50,15% ± 2,61%)
	5		168 (49,56% ± 2,61%)	161 (47,49% ± 2,60%)	152 (44,84% ± 2,59%)
N° Gausianas	1	Neutra	171 (50,44% ± 2,61%)	172 (50,74% ± 2,61%)	188 (55,46% ± 2,59%)
	2		136 (40,12% ± 2,56%)	159 (46,9% ± 2,60%)	185 (54,57% ± 2,60%)
	3		155 (45,72% ± 2,60%)	168 (49,56% ± 2,61%)	185 (54,57% ± 2,60%)
	4		157 (46,31% ± 2,60%)	168 (49,56% ± 2,61%)	177 (52,21% ± 2,60%)
	5		151 (44,54% ± 2,59)	165 (48,67% ± 2,61%)	159 (46,9% ± 2,60%)

Podemos observar como aumentan significativamente las tasas de identificación al entrenar y clasificar sólo con los ficheros de las emociones comunes en SES y EMODB, comparando los resultados obtenidos en la Tabla 57 y Tabla 58, con los obtenidos en la Tabla 52 y Tabla 53. De hecho, en este caso obtenemos tasas de identificación del orden de las que obteníamos cuando realizábamos los experimentos en los que entrenábamos y clasificábamos sólo ficheros de EMODB (ver Tabla 47*).*

Los vectores de entrenamiento y clasificación disponibles en estos experimentos son 220.995 y 93.876, respectivamente. Es decir, están aproximadamente en una proporción del 70% del total en entrenamiento y el 30% restante en clasificación, que consideramos porcentajes adecuados.

La mayor mejora que obtenemos al utilizar vectores normalizados, para 1, 3 y 4 gausianas, se obtiene normalizando con media y varianza estimada a partir de la voz neutra del locutor. Para 2 gausianas, se obtiene normalizando con media y varianza respecto a la voz del locutor. Y con 5 gausianas no mejoramos nada al normalizar, de

hecho, las tasas disminuyen. Analizamos a continuación la mejora obtenida al emplear vectores normalizados en función del número de gausianas empleado:

- *1 gausiana*: $\mathbf{\Delta E_{acc}}$ = (55,46% - 42,18%) = **13,28%**
- *2 gausianas*: $\mathbf{\Delta E_{acc}}$ = (58,11% - 42,48%) = **15,63%**
- *3 gausianas*: $\mathbf{\Delta E_{acc}}$ = (54,57% - 48,97%) = **5,6%**
- *4 gausianas*: $\mathbf{\Delta E_{acc}}$ = (52,21% - 49,26%) = **2,95%**

La mejora obtenida es elevada para el caso en el que utilizamos 1 y 2 gausianas, pero para el caso de 3 y 4 gausianas, la mejora disminuye considerablemente. De hecho, cuando utilizamos 4 gausianas, la mejora que hemos calculado no es estadísticamente fiable, ya que se solapan las bandas de fiabilidad de ambos resultados.

La mayor tasa de identificación (58,11%) se obtiene para el caso en el que utilizamos 2 gausianas con vectores normalizados respecto a la media y la varianza estimadas a partir de todos los datos de la voz del locutor.

En general, como veíamos antes, las mejores tasas de identificación se obtienen cuando utilizamos vectores normalizados con media y varianza, pero estas mejoras, comparándolas con las mejoras obtenidas al emplear vectores normalizados sólo respecto a la media o a la varianza, en muchos casos no son estadísticamente fiables. Así, en los únicos casos que no se solapan las bandas de fiabilidad es cuando empleamos 2 gausianas, estimando los vectores respecto a la voz del locutor y respecto a su voz neutra.

1.1.1.30. Análisis de la tasa de identificación para cada emoción

En este apartado vamos a analizar las tasas de identificación para cada emoción para los experimentos en los que entrenamos con datos de SES y clasificamos datos de EMODB, sólo con las emociones comunes a ambas bases de datos, en el caso en el que utilizamos 1 gausiana. Mostramos en distintas tablas los resultados obtenidos cuando utilizamos vectores sin normalizar (Tabla 59) y cuando utilizamos vectores normalizados respecto a la media y la varianza estimadas a partir de la voz del locutor (Tabla 60) o estimadas a partir de su voz neutra (Tabla 61). Hemos escogido este tipo de normalizaciones, dado que para ellas es para las que se obtiene la mayor mejora

utilizando 1 gausiana.

En las siguientes tablas (Tabla 59-*Tabla 61)* aparece, además de la tasa de identificación para cada emoción, la precisión de cada una de las emociones.

Tabla 59. Tasas de identificación para cada emoción del *experimento 20,* en el que entrenamos con SES y clasificamos con EMODB, con vectores sin normalizar, sólo con las emociones comunes.

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	98,59% (± 1,34%)	1,41%		
Enfado	96,85%	3,15% (± 1,49%)		
Tristeza	11,29%		85,48% (± 4,29%)	3,23%
Neutro	62,03%	3,80%	13,92%	20,25% (± 4,34%)
PRECISIÓN	36,68%	37,69%	85,99%	86,26%

Tabla 60. Tasas de identificación para cada emoción del *experimento 23,* en el que entrenamos con SES y clasificamos con EMODB, con vectores normalizados con media y varianza respecto a la voz del locutor, sólo con las emociones comunes.

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	77,46% (± 4,76%)	21,13%		1,41%
Enfado	77,17%	22,83% (± 3,58%)		
Tristeza			100%	
Neutro	12,66%	10,13%	41,77%	35,44% (± 5,17%)
PRECISIÓN	46,31%	42,22%	70,54%	96,18%

Tabla 61. Tasas de identificación para cada emoción del *experimento 26,* en el que entrenamos con SES y clasificamos con EMODB, con vectores normalizados con media y varianza respecto a la voz neutra del locutor, sólo con las emociones comunes.

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	74,65% (± 4,96%)	5,63%		19,72%
Enfado	79,53%	13,39% (± 2,9%)		7,09%
Tristeza			64,52% (± 5,83%)	35,48%
Neutro		1,27%		98,73% (± 1,21%)
PRECISIÓN	48,42%	65,99%	100%	61,32%

Observando las tablas, llegamos a las siguientes conclusiones:

- Las emociones que mejor se identifican son la alegría y la tristeza, por lo que podemos suponer que tanto la alegría como la tristeza comparten características comunes en ambos idiomas. La alegría tiene una tasa mayor al emplear vectores sin normalizar (98,59%) y la tristeza la conseguimos identificar siempre al emplear vectores normalizados respecto a la voz del locutor.
- En ambos casos, empleando vectores sin normalizar y normalizados, el enfado se confunde con la alegría con una tasa elevada, pero conseguimos reducir esta confusión al normalizar respecto a la voz del locutor (77,17% frente a 96,85%) y respecto a su voz neutra (79,53% frente a 96,85). Esto se puede deber a que el enfado representado en SES es un enfado en frío, mientras que el representado en EMODB es un enfado en caliente, y esto haga que se confunda con la alegría de SES (emoción con un nivel alto de activación).
- Normalizando respecto a la voz neutra del locutor conseguimos aumentar considerablemente la tasa de identificación de la neutra (de un 20,25% a un 98,73%), haciendo que no se confunda nunca con la alegría. Normalizando respecto a la voz del locutor lo que sucede es que la voz neutra deja de confundirse con la alegría y se confunde con la tristeza (41,77%). La confusión que se produce en la voz neutra nos puede hacer pensar que la forma de hablar en alemán es cercana a la forma en la que se interpretan las emociones

en castellano, y por ello se confunde tanto con la alegría como con la tristeza. Aunque hemos observado que somos capaces de reducir esta confusión al normalizar.

- Las confusiones que se producen del enfado con la alegría, hacen que a pesar de que la alegría sea la emoción para la que obtenemos la mayor tasa de identificación, su precisión sea bastante baja (36,68% al utilizar vectores sin normalizar y 46,31% y 48,2% al utilizar los normalizados). De hecho, obtenemos para el enfado, cuya tasa de identificación es muy baja, una precisión del mismo orden de la que obtenemos para la alegría e incluso mayor.
- La precisión de la tristeza es elevada, obteniéndose el 100% de precisión al emplear vectores normalizados respecto a la voz neutra del locutor.
- La precisión de la neutra es buena, a pesar de que sus tasas de identificación no lo son, debido a que la alegría y el enfado se confunden muy poco con ella (de hecho, cuando empleamos vectores sin normalizar, nunca se confunden con ella). Esta precisión disminuye considerablemente al normalizar respecto a la voz neutra del locutor (de 86,26% y 96,18% a 61,32%) debido a que la tristeza se confunde con ella cuando no se identifica.
- La alegría y el enfado nunca se confunden con la tristeza.
- La tristeza nunca se confunde con el enfado.

6.3.3.Entrenamiento con datos de EMODB y clasificación de datos de SES, utilizando sólo las emociones comunes

En este apartado realizaremos los experimentos de identificación de emociones en los que extraeremos los modelos a partir de los datos de las emociones comunes de EMODB y a continuación clasificaremos los ficheros de SES, sólo los de las emociones comunes. De esta forma evaluaremos la capacidad que nos ofrecen los modelos obtenidos a partir de las emociones en alemán, para identificar las emociones en castellano.

1.1.1.31.Descripción de los experimentos.

Los ficheros de SES que clasifiquemos con los modelos obtenidos a partir de los ficheros de EMODB, pueden ser los párrafos, las frases o ambos. Hemos realizado los distintos experimentos, obteniendo los mejores resultados medios cuando clasificamos sólo los párrafos, así que estos resultados serán los que analicemos en el siguiente apartado.

Los posibles experimentos que obtenemos según normalicemos o no los vectores de características, y en función de la normalización aplicada en su caso, son los que se enuncian a continuación:

- EXPERIMENTO 27: *sin normalizar los vectores de características.*
- EXPERIMENTO 28: *CMN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 29: *CVN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 30: *CMN + CVN (normalizando respecto a la voz del locutor).*
- EXPERIMENTO 31: *CMN (normalizando respecto a la voz neutra del locutor).*
- EXPERIMENTO 32: *CVN (normalizando respecto a la voz neutra del locutor).*
- EXPERIMENTO 33: *CMN + CVN (normalizando respecto a la voz neutra del locutor).*

1.1.1.32.Resultados de los experimentos

En las siguientes tablas (Tabla 62 y *Tabla 63*) se muestran el número de ficheros identificados, así como la tasa de identificación y su correspondiente banda de fiabilidad, para cada uno de los experimentos de identificación en los que entrenamos con datos de las emociones comunes de EMODB y clasificamos los párrafos de las emociones comunes de SES. Se muestran en diferentes tablas los resultados obtenidos para el experimento en el que utilizamos vectores sin normalizar (*experimento 27)* y los resultados para los experimentos con vectores normalizados (*experimentos 28-33).*

El número total de ficheros de clasificación que tenemos en todos los experimentos de este apartado, teniendo en cuenta que sólo hemos utilizado para la clasificación los párrafos de SES, por ser éstos con los que obtenemos los mejores

resultados medios (como comentábamos en la descripción de los experimentos), es **44**.

Tabla 62. Número de ficheros identificados, tasa de identificación y banda de fiabilidad del *experimento 27,* en el que entrenamos con datos de EMODB y clasificamos datos de SES, sin normalizar, sólo con las emociones comunes, en función del número de gausianas.

		Sin normalizar
N° Gausianas	1	20 (45,45% ± 7,21%)
	2	20 (45,45% ± 7,21%)
	3	21 (47,73% ± 7,23%)
	4	29 (65,91% ± 6,86%)
	5	29 (65,91% ± 6,86%)

Tabla 63. Número de ficheros identificados, tasa de identificación y banda de fiabilidad de los *experimentos 28-33*, en los que entrenamos con datos de EMODB y clasificamos datos de SES, normalizando, sólo con las emociones comunes, en función del número de gausianas.

		Normalizado respecto a...	Media (μ)	Varianza (σ^2)	Media (μ) + Varianza (σ^2)
N° Gausianas	1	Locutor	15 (34,09% ± 6,86%)	26 (59,09% ± 7,12%)	21 (47,73% ± 7,23%)
	2		12 (27,27% ± 6,45%)	23 (52,27% ± 7,23%)	15 (34,09% ± 6,86%)
	3		13 (29,55% ± 6,6%)	22 (50% ± 7,24%)	14 (31,82% ± 6,74%)
	4		17 (38,64% ± 7,05%)	26 (59,09% ± 7,12%)	15 (34,09% ± 6,86%)
	5		17 (38,79% ± 7,05%)	26 (59,09% ± 7,12%)	14 (31,82% ± 6,74%)
N° Gausianas	1	Neutra	17 (38,64% ± 7,05%)	21 (47,73% ± 7,23%)	17 (38,64% ± 7,05%)
	2		13 (29,55% ± 6,6%)	20 (45,45% ± 7,21%)	8 (18,18% ± 5,58%)
	3		16 (36,36% ± 6,96%)	18 (40,91% ± 7,12%)	10 (22,73% ± 6,07%)
	4		20 (45,45% ± 7,21%)	19 (43,18% ± 7,17%)	20 (45,45% ± 7,21%)
	5		13 (29,55% ± 6,6%)	29 (65,91% ± 6,86%)	15 (34,09% ± 6,86%)

Observando los resultados mostrados en las tablas vemos que las tasas de identificación obtenidas al emplear vectores normalizados, no son muy superiores a las obtenidas cuando empleamos vectores sin normalizar. De hecho, sólo obtenemos mejoras cuando utilizamos 1, 2 ó 3 gausianas con vectores normalizados con la varianza estimada a partir de los datos del locutor, y dada la amplitud de las bandas de fiabilidad, estás no son fiables, ya que las mejoras obtenidas son:

- *1 gausiana*: $\mathbf{\Delta E_{acc}}$ = (59,09% - 45,45%) = **13,64%**

  Banda de fiabilidad = 7,21% + 7,12% = 14,33%

- *2 gausianas*: $\mathbf{\Delta E_{acc}}$ = (52,27% - 45,45%) = **6,82%**

  Banda de fiabilidad = 7,21% + 7,23% = 14,44%

- *3 gausianas*: $\mathbf{\Delta E_{acc}}$ = (50% - 47,73%) = **2,27%**

  Banda de fiabilidad = 7,23% + 7,24% = 14,47%

Observamos como la mejora va disminuyendo a medida que aumenta el número de gausianas, mientras que lo lógico sería que sucediera lo contrario.

Para estos experimentos, el número de vectores de entrenamiento es menor que el número de vectores de clasificación (93.876 frente a 187.452), lo que hace que obtengamos modelos para cada emoción basados en pocos datos.

La mayor tasa de identificación (65,91%) se obtiene cuando utilizamos vectores sin normalizar con 4 y con 5 gausianas, y utilizando vectores normalizados respecto a la varianza estimada a partir de la neutra del locutor, con 5 gausianas.

1.1.1.33. Análisis de la tasa de identificación para cada emoción

En este apartado vamos a analizar las tasas de identificación para cada emoción para los experimentos en los que entrenamos con datos EMODB y clasificamos datos de SES, sólo con las emociones comunes, en el caso en el que utilizamos 5 gausianas. Mostramos en distintas tablas los resultados obtenidos cuando utilizamos vectores sin normalizar (Tabla 64) y cuando utilizamos vectores normalizados respecto a la varianza estimada a partir de la voz neutra del locutor (Tabla 65). La razón por la que mostramos los datos relativos a este tipo de normalización y con este número de gausianas, es porque en el apartado anterior hemos comprobado que era para la que se obtenían la mayor tasa de identificación media.

En las siguientes tablas (Tabla 64 y Tabla 65) aparece, además de a tasa de identificación para cada emoción, la precisión de cada una de las emociones.

Tabla 64. Tasas de identificación para cada emoción del *experimento 27,* en el que entrenamos con EMODB y clasificamos los párrafos de SES, con vectores sin normalizar, sólo con las emociones comunes, utilizando 5 gausianas.

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	75% (± 12%)			25%
Enfado	66,67			33,33%
Tristeza			100%	
Neutro				100%
PRECISIÓN	53%		100%	63%

Tabla 65. Tasas de identificación para cada emoción del *experimento 32,* en el que entrenamos con EMODB y clasificamos los párrafos de SES, con vectores normalizados respecto a la varianza estimada a partir de la voz neutra del locutor, sólo con las emociones comunes, utilizando 5 gausianas.

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	66,67 (± 13,06)%	8,33%		25%
Enfado	58,33%	8,33% (± 7,66%)		33,33%
Tristeza			100%	
Neutro				100%
PRECISIÓN	53%	50%	100%	63%

Los resultados obtenidos en empleando vectores sin normalizar o normalizados son similares:

- La tristeza y la neutra siempre se identifican. Adicionalmente, ninguna de las emociones se confunden con la tristeza, por lo que ésta tiene una precisión del 100%.
- La alegría tiene una tasa mayor al emplear vectores sin normalizar (75% frente a 66,67%), confundiéndose con la neutra cuando no se identifica.
- El enfado se confunde con una tasa elevada con la alegría y también con la neutra. No se identifica nunca, en el caso de emplear vectores sin normalizar, y sólo con un 8,33%, al emplear vectores normalizados.

- Debido a las confusiones que se producen de la alegría y el enfado con la neutra, la precisión de ésta es baja (63%), a pesar de que siempre se identifica.

Dados los curiosos resultados obtenidos al analizar la tasa de identificación de las distintas emociones, vamos a analizar las tasas en el caso en el que en vez de clasificar los párrafos de SES, clasificamos las frases grabadas de forma independiente. Las siguientes tablas muestran los resultados obtenidos cuando utilizamos vectores sin normalizar (Tabla 66) y cuando utilizamos vectores normalizados respecto a la varianza estimada a partir de la voz neutra del locutor (Tabla 67*),* utilizando en ambos casos 5 gausianas.

Tabla 66. Tasas de identificación para cada emoción del *experimento 27,* en el que entrenamos con EMODB y clasificamos las frases de SES, con vectores sin normalizar, sólo con las emociones comunes, utilizando 5 gausiana

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	53,33% (± 7,14%)	2,22%		44,44%
Enfado	48,89%	31,11%(± 6,63%)		20%
Tristeza			28,89% (± 6,49%)	71,11%
Neutro				100%
PRECISIÓN	52%	93%	100%	42%

Tabla 67. Tasas de identificación para cada emoción del *experimento 32,* en el que entrenamos con EMODB y clasificamos los párrafos de SES, con vectores normalizados respecto a la varianza estimada a partir de la voz neutra del locutor, sólo con las emociones comunes, utilizando 5 gausiana

	EMOCIÓN IDENTIFICADA			
EMOCIÓN INTERPRETADA	Alegría	Enfado	Tristeza	Neutro
Alegría	31,11% (± 6,63%)	8,89%		60%
Enfado	24,44%	28,89% (± 6,49%)		46,67%
Tristeza			42,22% (± 7,07%)	57,78%
Neutro				100%
PRECISIÓN	56%	76%	100%	38%

Las conclusiones que obtenemos observando estas tablas son las siguientes:

- La neutra se sigue interpretando siempre, aunque en este caso, el resto de emociones se confunden mucho con ella, por lo que su precisión es bastante baja (42% y 38%).

- La tristeza en este caso no se identifica siempre, sino que se confunde con la neutra (con un 71,11% al emplear vectores sin normalizar y 57,78% al emplear vectores normalizados).

- La tasa de identificación de la alegría se ve reducida respecto a los experimentos en los que clasificábamos los párrafos. Sin embargo, en este caso el enfado se identifica, aunque su tasa es baja (31,11% y 28,89%, con vectores sin o con normalización, respectivamente).

6.3.4. Conclusiones de los experimentos de identificación con distintos idiomas con características basadas en MFCC

Una vez analizados los resultados obtenidos en los distintos experimentos de identificación en los que entrenamos con los datos de una de las bases de datos y clasificamos los de la otra, las conclusiones a las que llegamos son:

- *Entrenando con SES y clasificando con EMODB con todas las emociones*:
    - No obtenemos una mejora muy elevada al normalizar. El mayor valor que obtenemos de mejora es un 7,29%, obtenido en el caso de utilizar 2 gausianas, cuando normalizamos con media y varianza respecto a la voz neutra del locutor.
    - Normalizando los vectores de características sólo respecto a la media, obtenemos tasas de identificación menores que cuando no normalizamos.
    - La emoción que mejor se identifica es la tristeza, identificándose siempre al emplear vectores normalizados respecto a la voz del locutor.

La precisión más elevada también la obtiene la tristeza.

- Normalizando con la media y la varianza respecto a la voz neutra del locutor conseguimos aumentar la tasa de identificación de la neutra de un 25,15% a un 98,73%.
- Todas las emociones, menos la tristeza, se confunden con la sorpresa. Al normalizar conseguimos reducir la tasa de confusión de todas las emociones con la sorpresa, pero algunas se confunden con otras:
  - El aburrimiento se confunde con la tristeza y la neutra.
  - El asco se confunde con la alegría y la neutra.
  - Y la neutra se confunde con la tristeza.
- Excepto para la neutra, las precisiones más elevadas se obtienen cuando empleamos vectores normalizados respecto a la voz neutra del locutor.
- Se cumple que la alegría y el enfado nunca se confunden con la tristeza, y la tristeza nunca se confunde con el enfado.

- *Entrenando con SES y clasificando con EMODB sólo con las emociones comunes:*
  - Obtenemos tasas de identificación medias del orden de las que obteníamos cuando entrenábamos y clasificábamos sólo con datos de EMODB.
  - La normalización que mejor funciona es en la que estimamos la media y la varianza a partir de los datos de la voz neutra del locutor. La mayor mejora se consigue con 2 gausianas (15,63%) y también con 2 gausianas es con las que conseguimos la mayor tasa de identificación (58,11%).
  - La emoción que mayor tasa de identificación tiene es la alegría, seguida de la tristeza. Estas tasas son mayores cuando utilizamos vectores sin normalizar para la alegría (98,59%) y cuando normalizamos con media y varianza respecto a la voz del locutor para la tristeza (100%).

  - El enfado se confunde con la alegría, lo que hace que ésta tenga una baja precisión aunque su tasa de identificación sea elevada.
  - Al emplear vectores normalizados, conseguimos reducir la tasa de confusión del enfado y principalmente de la neutra, aumentando considerablemente la tasa de identificación de la neutra.
  - Teniendo en cuenta la tasa de identificación y la precisión, la emoción que mejor se reconoce es la tristeza, ya que casi es la que menos se confunde con otras emociones y adicionalmente, cuando una emoción se identifica como tristeza, es bastante probable que lo sea.
  - Se cumple que la alegría y el enfado nunca se confunden con la tristeza.

- *Entrenando con EMODB y clasificando con SES sólo con las emociones comunes:*
  - En este tipo de experimentos, la normalización para la que obtenemos mejores resultados es aquella en la que estimamos sólo la varianza a partir de todos los datos de la voz del locutor. Recordamos que normalizando con la varianza, se intentan compensar las características propias del locutor. Pero las mejoras obtenidas no son significativas.
  - Clasificando los párrafos de SES, la tristeza y la neutra siempre se identifican. La alegría tiene una elevada tasa de identificación (75% si empleamos vectores sin normalizar y 66% si empleamos vectores normalizados respecto a la varianza del locutor). El enfado no se identifica y se confunde, principalmente, con la alegría. Ninguna emoción se confunde con la tristeza, por lo que su precisión es 100%.
  - Clasificando las frases de SES, la neutra también se identifica siempre, sin embargo, la tristeza se confunde con la neutra. Al igual que cuando clasificamos los párrafos, el enfado se confunde con la alegría, y la alegría se confunde con la neutra (pero en este caso, con una tasa de confusión mayor).

- De los tres tipos de experimentos que hemos analizado, las conclusiones generales que obtenemos son que la alegría y la tristeza parece que

comparten características comunes en ambos idiomas, ya que al entrenar con datos de SES y clasificar los de EMODB, son las que mejor se reconocen. Es importante el hecho de que entrenando con datos de SES y clasificando los de EMODB, conseguimos una tasa de identificación de la alegría, mayor de la que conseguíamos cuando trabajábamos sólo con datos de EMODB. Pero su precisión es baja debido a que el enfado se confunde con ella con una elevada tasa de confusión, pudiendo deberse este hecho a que el enfado interpretado en EMODB es un enfado en caliente que puede tener ciertas similitudes con la alegría interpretada por el actor de SES.

6.4.Conclusiones de los experimentos de identificación con características basadas en MFCC

Una vez analizados los resultados obtenidos en los experimentos utilizando las características segmentales, las conclusiones generales que podemos extraer son las que se exponen a continuación. Primero mostraremos las conclusiones comunes en todos los experimentos realizados en este capítulo. A continuación expondremos las conclusiones particulares extraídas de los experimentos realizados con SES, con EMODB y con ambas bases de datos.

- La normalización con la que normalmente conseguimos mejores resultados es aquella en la que estimamos la media y la varianza, ya sea respecto a la voz del locutor o respecto a su voz neutra. Como excepción tenemos el caso en el que entrenamos con datos de EMODB y clasificamos datos de SES, siendo la mejor normalización aquella en la que sólo utilizamos la varianza estimada a partir de todos los datos de la voz del locutor.

- Las mejoras obtenidas con EMODB y en los experimentos con distintos idiomas son mayores, en valor absoluto, que las obtenidas con SES (las mejoras son: 7,39% para SES, 15,51% para EMODB y 15,63% para los experimentos con distintos idiomas). Sin embargo, la mejora del error relativo es mayor en el caso de SES, debido a que la tasa de identificación inicial es mayor que la de los otros dos casos.

- La emoción que mejor se reconoce en todos los experimentos realizados es la **tristeza**, teniendo en cuenta tanto la tasa de identificación como la precisión. Es decir, no sólo es que se reconozca, sino que el resto de emociones no se confunden con ella, lo que nos hace suponer que tiene un patrón muy diferente al resto de emociones.

- Tanto en SES como en EMODB, la emoción que peor se identifica es la **alegría**. Sin embargo, cuando entrenamos con datos de SES y clasificamos datos de EMODB (sólo con las emociones comunes), obtenemos una elevada tasa de identificación de la alegría, pero el enfado y la neutra se confunden con

ella, lo que hace que su precisión sea baja. En general, la emoción que peor se identifica en los experimentos con distintos idiomas es el enfado, que se confunde tanto con la alegría como con la neutra.

- La neutra es la emoción que tiene un mayor aumento de la tasa de identificación al emplear vectores de características normalizados, particularmente cuando normalizamos con media y varianza estimadas respecto a la voz neutra del locutor.

- En general, al normalizar aumenta la precisión de la mayoría de las emociones, pero esto no es así al realizar los experimentos con distintos idiomas.

- En todos los experimentos realizados con características basadas en MFCC se cumple que el enfado nunca se confunde con la tristeza, por lo que podemos concluir que las características segmentales de estas emociones son muy diferentes. En general, la alegría tampoco suele confundirse con la tristeza, ni la tristeza con la alegría.

- En los experimentos realizados con datos de SES, se obtienen mejores resultados cuando empleamos como unidad de clasificación el párrafo, debido a que a la hora de tomar la decisión de qué emoción se trata, lo hacemos sobre un mayor número de vectores de características. En general, aumenta la tasa de identificación al aumentar el número de gausianas. En estos experimentos, la sorpresa no se identifica con una elevada tasa (se confunde con la alegría, y la alegría se confunde con la sorpresa), pero conseguimos que mejore su tasa de identificación al normalizar.

- En los experimentos realizados con datos de EMODB, la emoción que peor se identifica (junto con la alegría) es el miedo. Al normalizar, conseguimos mejorar la tasa de identificación del asco con un 34,78% de mejora.

- En los experimentos en los que entrenamos con datos de SES y clasificamos datos de EMODB, con todas las emociones, se confunden todas ellas con la sorpresa, a excepción de la tristeza. Al normalizar, conseguimos reducir esa confusión, pero sin embargo, el aburrimiento se confunde con la tristeza y con

la neutra; el asco con la alegría y la neutra; y la neutra con la tristeza.

- En los experimentos en los que entrenamos con datos de SES y clasificamos datos de EMODB, sólo con las emociones comunes, obtenemos tasas de identificación medias del orden de las que obtenemos en el caso en el que sólo usamos datos de EMODB.

- En los experimentos en los que entrenamos con datos de EMODB y clasificamos datos de SES, sólo con las emociones comunes, la neutra siempre se identifica, sin embargo, la tristeza se identifica siempre si clasificamos los párrafos, pero se confunde con la neutra si clasificamos las frases independientes. La alegría tiene una elevada tasa al clasificar los párrafos, que se ve reducida al clasificar las frases, debido a que se confunde con la neutra. El enfado se confunde principalmente con la alegría, y también con la neutra.

7. IDENTIFICACIÓN DE EMOCIONES BASADA EN INFORMACIÓN PROSÓDICA

En este capítulo se definen y analizan los resultados obtenidos en los experimentos de identificación de emociones basados en información prosódica. Estos experimentos los hemos realizado sólo con la base de datos SES, porque en ella disponemos de la información lingüística necesaria para poder dividir las frases en grupos fónicos, que serán de los que extraeremos las características empleadas en el entrenamiento y la clasificación. La forma en la que dividimos los ficheros en los distintos grupos fónicos, fue explicada con detalle en el *capítulo 1.1.1.17*.

Dentro de estos experimentos podemos distinguir cuatro tipos en función de las características que contengan los vectores de entrenamiento y clasificación. En dos de ellos, las características están relacionadas con la frecuencia fundamental, la diferencia está en que en uno incluimos su valor medio y en el otro no. En los otros dos experimentos, las características están relacionadas con el ritmo, bien de toda la locución o bien de cada uno de los grupos fónicos.

Lo primero que vamos a analizar son los modelos obtenidos en el entrenamiento para las características prosódicas que vamos a emplear en los experimentos, de forma que podamos intuir cuales serán las características más relevantes.

A continuación describiremos los distintos experimentos realizados. Cada uno de ellos lo realizaremos entrenando con dos de las sesiones de cada una de las emociones y una sesión de la neutra, y clasificando la sesión que nos queda de cada emoción y de la neutra. De esta forma obtenemos tres experimentos diferentes para cada tipo de experimento descrito.

En la realización de estos experimentos sólo utilizamos una gausiana, debido a que no es aconsejable utilizar más, dada la escasez de datos para el entrenamiento.

7.1.Modelos obtenidos para las diferentes características prosódicas

En este apartado analizaremos, mediante la representación de gausianas, los modelos obtenidos en el entrenamiento para las diferentes características prosódicas vistas con detalle en el *apartado 1.1.1.17*, que son las siguientes:

- Valor medio de la frecuencia fundamental (F0).
- Mínimo de F0.
- Máximo de F0.
- Rango de F0.
- Pendiente de subida.
- Pendiente de bajada.
- Velocidad de locución de la frase.
- Velocidad de locución de cada grupo fónico.

Para cada una de las características representaremos el modelo obtenido en el entrenamiento para los cuatro posibles casos que hemos considerado: utilizando todos los grupos fónicos, utilizando sólo los grupos fónicos iniciales, utilizando sólo los grupos fónicos finales y utilizando sólo los grupos fónicos medios.

En todas las representaciones que se muestran a continuación, la leyenda de los colores de las gráficas es la siguiente:

- Alegría: azul oscuro.
- Enfado: rosa.
- Sorpresa: verde.
- Tristeza: azul claro.
- Neutro: morado.

7.1.1.Valor medio de F0

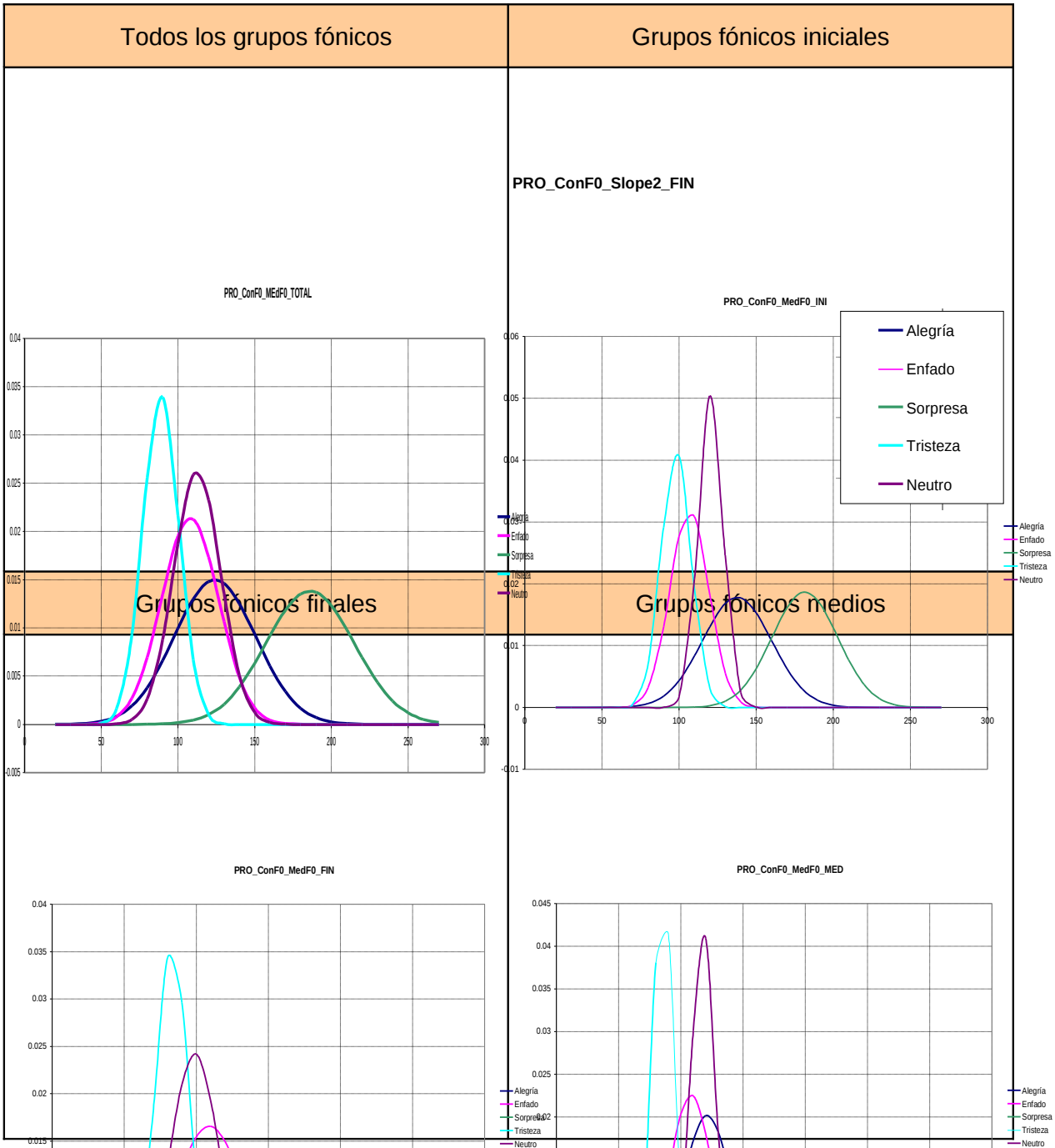
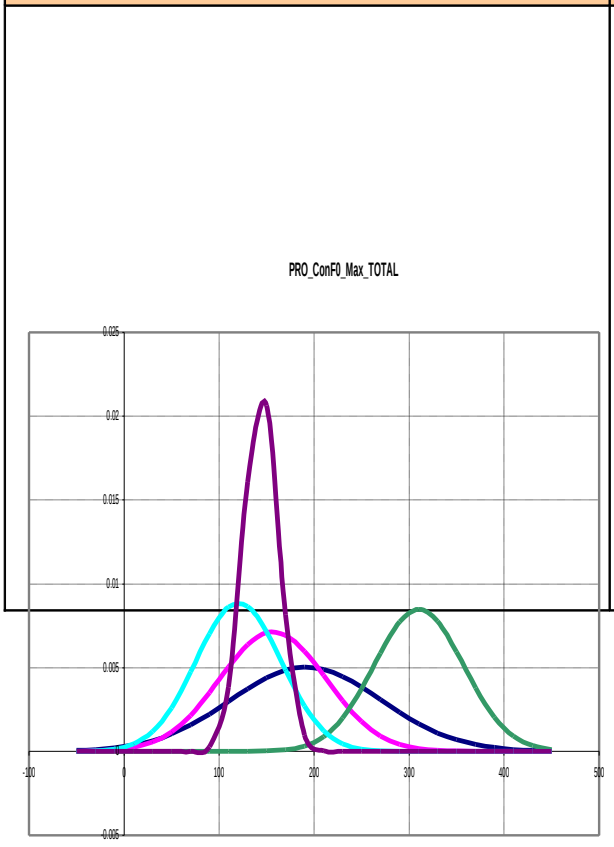
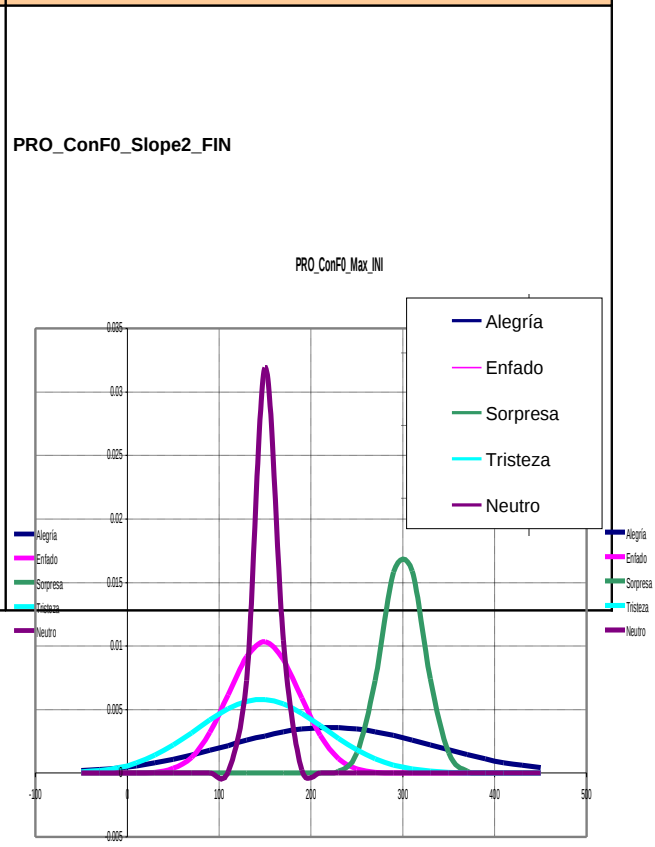


Figura 8. Modelos obtenidos para el valor medio de F0.

Observando los modelos obtenidos para el valor medio de F0 vemos que el que aparece más diferenciado es el de la sorpresa, y en algunas ocasiones también el de la tristeza y la neutra. Sin embargo, la alegría y el enfado presentan modelos poco diferenciados del resto. Las conclusiones que sacamos observando estas gráficas son:

- El valor medio de la sorpresa es más elevado que el del resto de emociones, por lo que permite que se identifique mejor esta emoción. El caso en el que la sorpresa presenta un modelo más diferenciado es cuando entrenamos solamente con los grupos fónicos medios.
- En el caso de utilizar los grupos fónicos iniciales y los medios, la tristeza y la neutra, también aparecen diferenciadas del resto, lo que nos permitirá su mejor identificación.
- Cuando utilizamos los grupos fónicos finales, se solapan las gausianas de unas emociones con otras, por lo que en este caso, lo que hacemos es introducir confusión.

7.1.2.Valor máximo de F0

Todos los grupos fónicos	Grupos fónicos iniciales
	

Grupos fónicos finales	Grupos fónicos medios

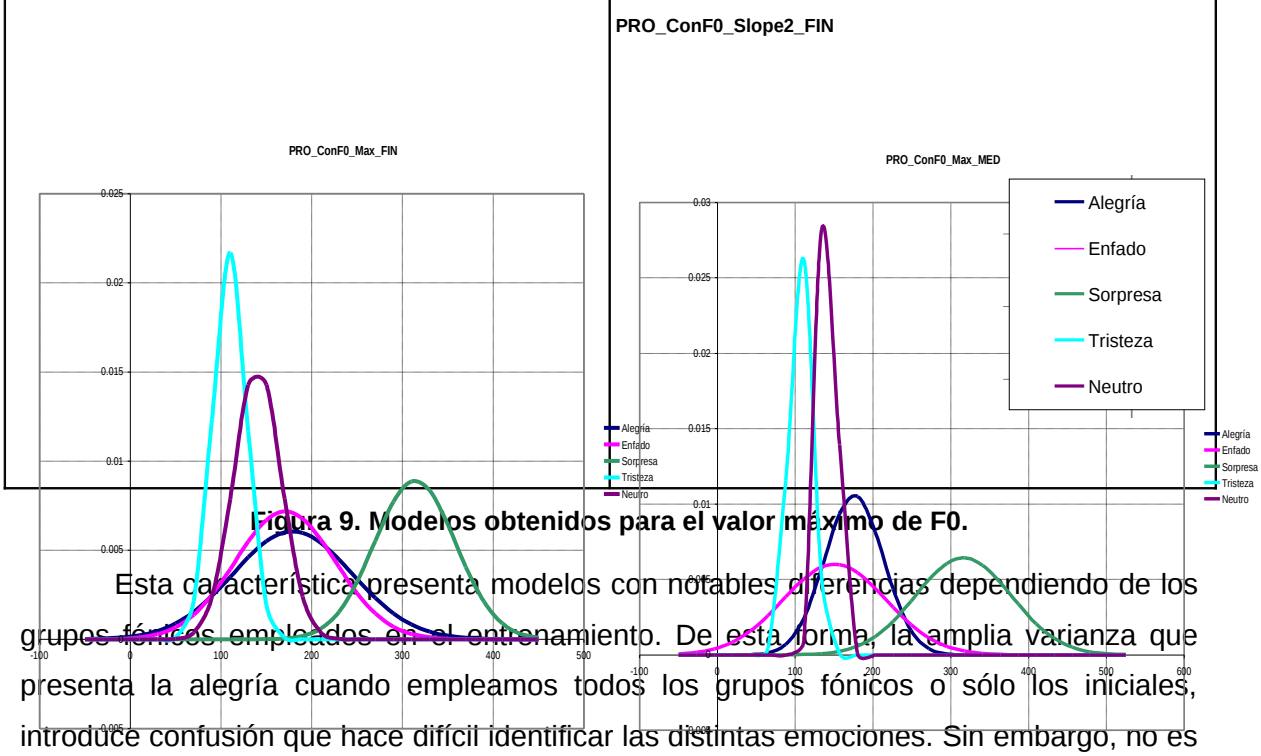


Figura 9. Modelos obtenidos para el valor máximo de F0.

Esta característica presenta modelos con notables diferencias dependiendo de los grupos fónicos empleados en el entrenamiento. De esta forma, la amplia varianza que presenta la alegría cuando empleamos todos los grupos fónicos o sólo los iniciales, introduce confusión que hace difícil identificar las distintas emociones. Sin embargo, no es así, para los grupos fónicos finales y medios, donde la confusión introducida es menor. Las conclusiones que sacamos observando las gráficas son:

- Al igual que en el caso anterior, la sorpresa es la que aparece más diferenciada del resto de emociones en todos lo casos.
- Cuando solamente utilizamos los grupos fónicos iniciales, se introduce confusión, ya que la media del enfado, la tristeza y la neutra es prácticamente la misma, y la alegría tiene una elevada varianza, interfiriendo de esta forma con la sorpresa que aparece desplazada.
- La tristeza y la voz neutra aparecen más diferenciadas en el caso en el que entrenamos con los grupos fónicos medios.
- La voz neutra presenta muy poca varianza del valor máximo de F0, salvo en el caso en el que utilizamos los grupos fónicos finales.

- La varianza del valor máximo de F0 que presenta la alegría es elevado, principalmente cuando empleamos todos los grupos fónicos y sólo los iniciales.
- El modelo obtenido para la alegría y el enfado cuando usamos los grupos fónicos finales es muy similar.

7.1.3.Valor mínimo de F0

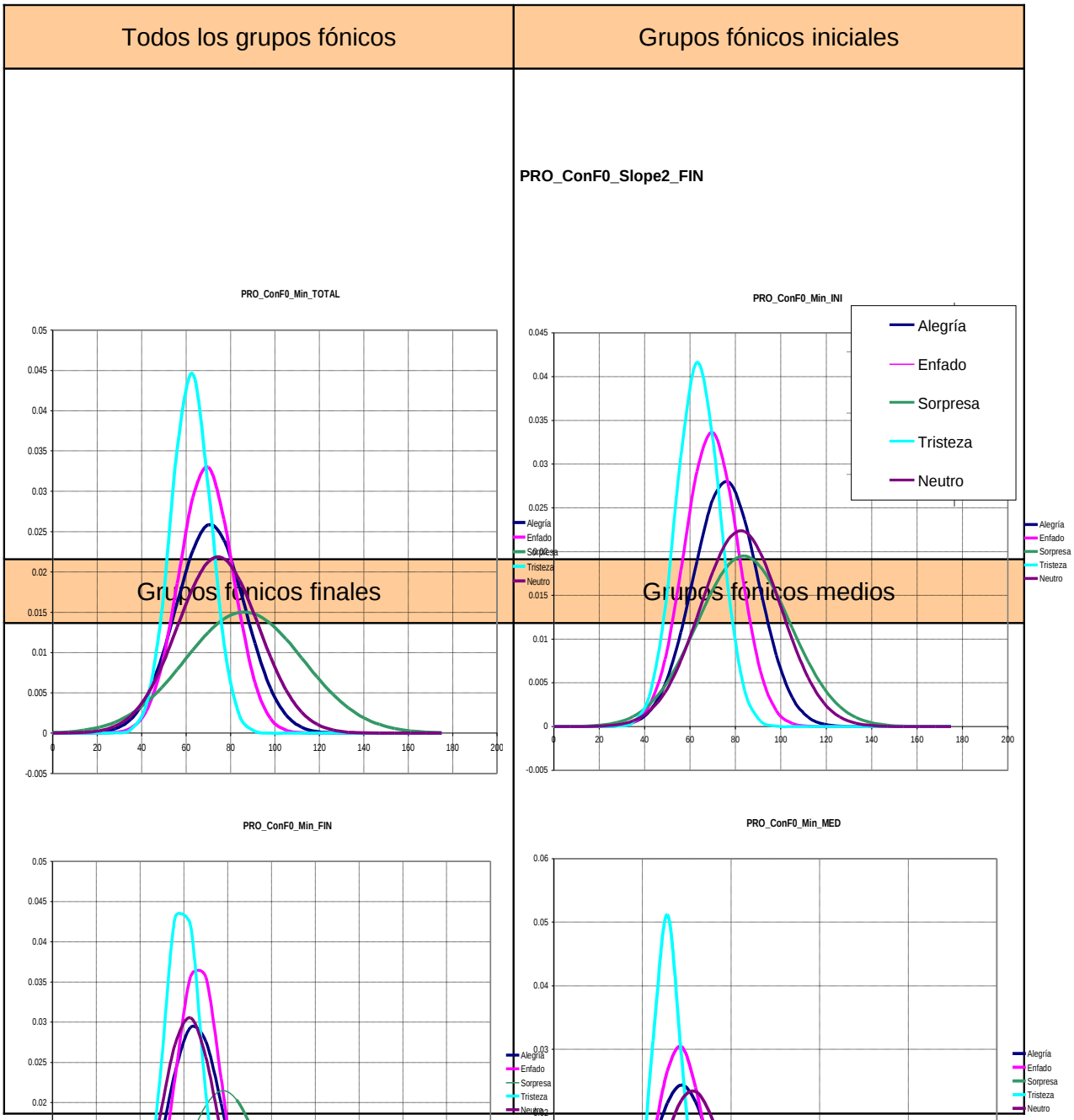


Figura 10. Modelos obtenidos para el valor mínimo de F0.

Esta característica no nos aporta prácticamente ninguna información acerca de qué emoción puede ser la que vamos a clasificar, ya que los modelos que obtenemos se solapan. Particularmente, los modelos obtenidos cuando usamos los grupos fónicos finales y medios, para la alegría y el enfado son muy similares.

7.1.4.Rango de F0

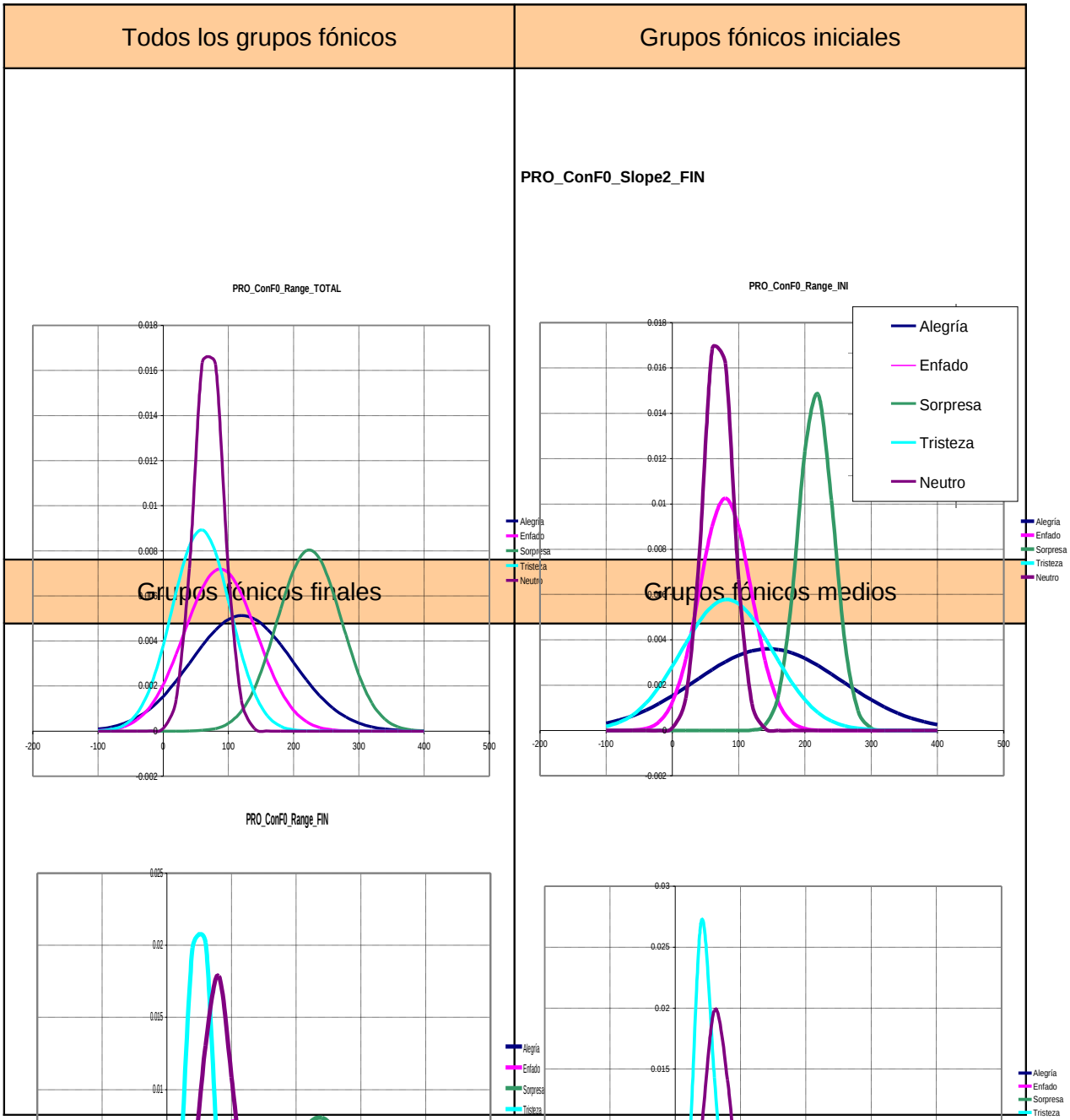
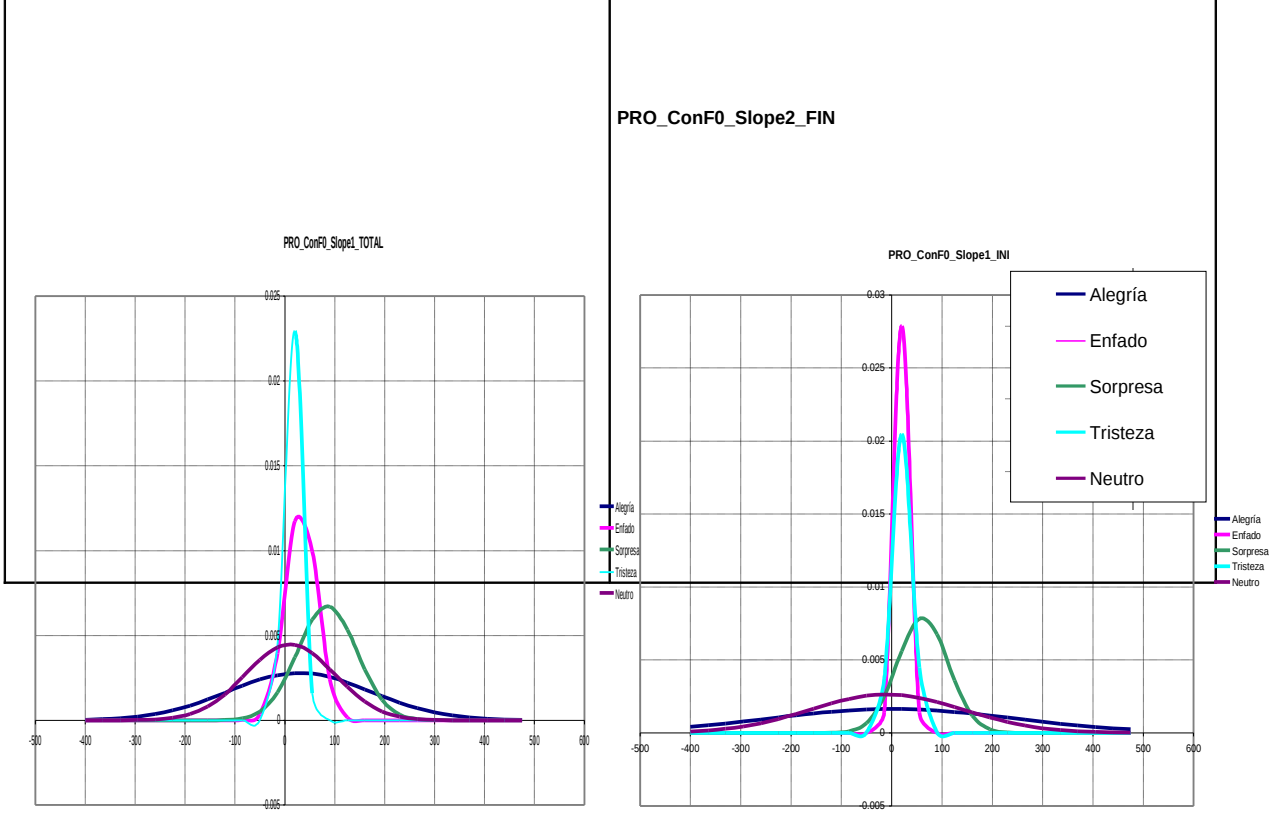


Figura 11. Modelos obtenidos para el rango de F0.

Observando los modelos obtenidos para el rango de F0 que presentan los distintos grupos fónicos llegamos a las siguientes conclusiones:

- Al igual que en las dos primeras características, la gausiana que aparece diferenciada es la correspondiente a la sorpresa.
- Cuando usamos los grupos fónicos iniciales se produce confusión entre unas gausianas y otras, ya que la media del enfado, la tristeza y la neutra es similar, y la alegría tiene una amplia varianza, interfiriendo así con la sorpresa.
- La varianza del rango de F0 de la voz neutra es pequeña.
- El modelo obtenido para la alegría y el enfado cuando usamos los grupos fónicos finales es muy similar.

7.1.5.Pendiente de subida

Todos los grupos fónicos	Grupos fónicos iniciales
	

PRO_ConF0_Slope2_FIN

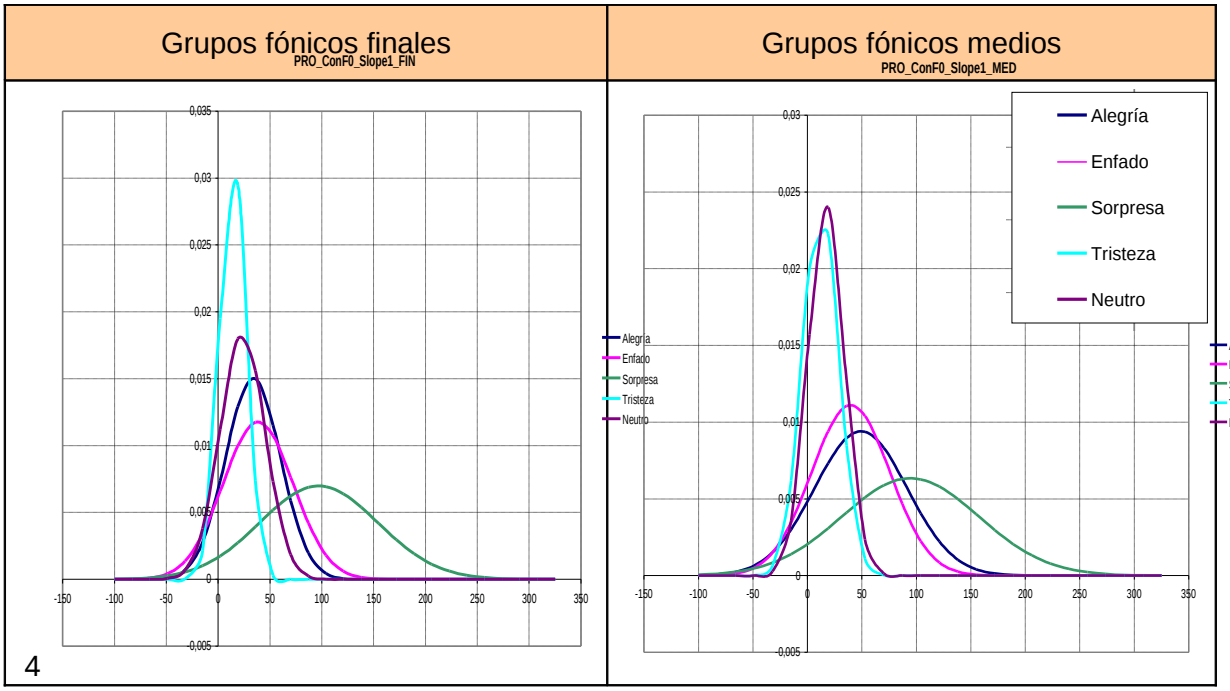


Figura 12. Modelos obtenidos para la pendiente de subida.

Los modelos obtenidos para la pendiente de subida no nos van a permitir identificar las distintas emociones, dado que se solapan unos con otros. Vemos a continuación las conclusiones extraídas de los dibujos:

- En este caso, la sorpresa no aparece tan diferenciada del resto de emociones.
- Cuando usamos los grupos fónicos iniciales se introduce confusión, ya que la media del enfado y la tristeza es prácticamente la misma, y la alegría y la neutra tienen una varianza muy elevada, interfiriendo con todas las demás. Cuando utilizamos todos los grupos fónicos también pasa esto, pero la varianza de la neutra es menor que en el caso de utilizar sólo los grupos fónicos iniciales.
- Cuando usamos los grupos fónicos finales también se introduce confusión, ya que aunque la varianza de las emociones no sea tan elevada como en el caso anterior, todas presentan una media muy próxima, lo que hace que se solapen los distintos modelos.
- Cuando entrenamos con todos los grupos fónicos y sólo con los iniciales, la varianza de la alegría y el enfado es muy elevada.
- La varianza de la tristeza es pequeña.

7.1.6.Pendiente de bajada

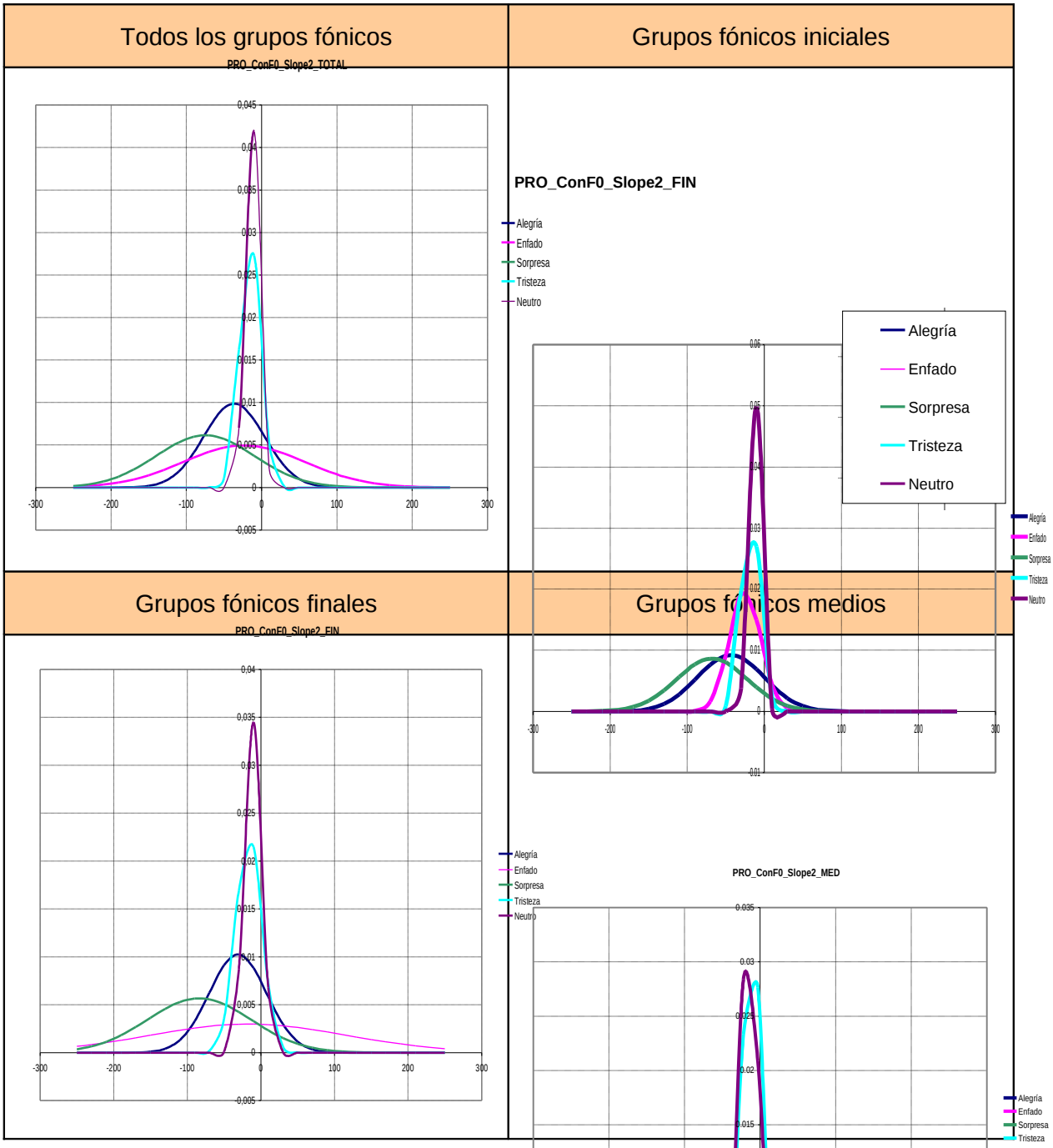


Figura 13. Modelos obtenidos para la pendiente de bajada.

Al igual que pasaba con la pendiente de subida, la pendiente de bajada no presenta unos modelos diferenciados para las emociones, que nos permitan identificarlas. Las principales conclusiones que obtenemos observando las figuras son las siguientes:

- La varianza del enfado tanto en el caso en el que utilizamos todos los grupos fónicos, como cuando sólo usamos los finales, es elevada, interfiriendo esta emoción con el resto.
- En los cuatro casos (todos los grupos fónicos, sólo los iniciales, sólo los finales y sólo los medios) los modelos obtenidos para la tristeza y la neutra son similares.
- La varianza de la voz neutra es pequeña.
- Cuando entrenamos con los grupos fónicos medios, los modelos obtenidos para la voz neutra y la tristeza son similares. Y los obtenidos para la alegría y el enfado, en este caso, también.

7.1.7.Velocidad de locución de la frase

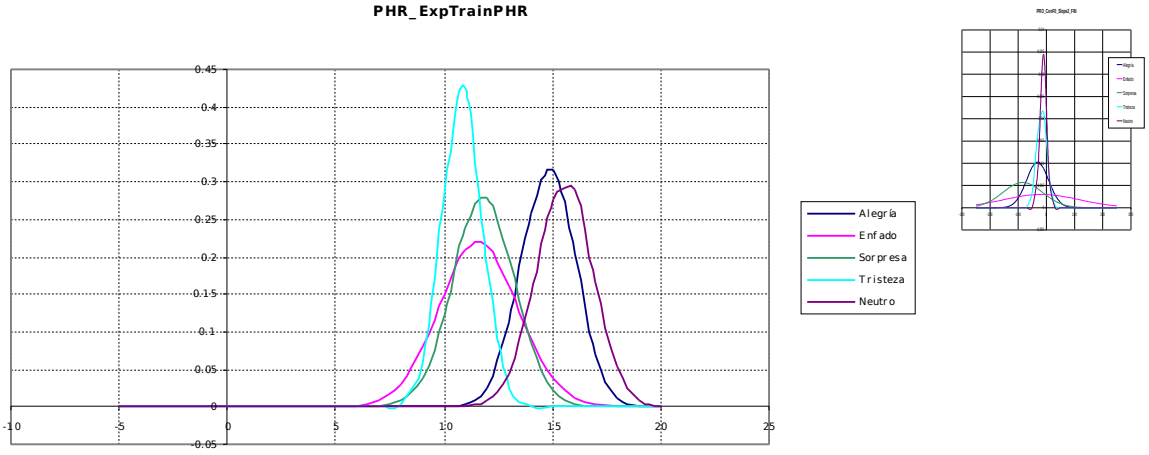


Figura 14. Modelos obtenidos para la velocidad de locución de la frase.

- En este caso podemos observar que las emociones aparecer separadas como en dos grupos: por un lado el enfado, la sorpresa y la tristeza, y por el otro, la alegría y la neutra. Pero dentro de estos grupos, los modelos son similares, por lo que va a ser complicado identificar las distintas emociones sólo con esta característica.
- Las emociones que más alejadas aparecen son la tristeza y la neutra, lo que conlleva que son las emociones que mejor se van a distinguir entre ellas, es decir, para identificar si se trata de una frase triste o neutra, la velocidad de locución de la frase nos va a ayudar bastante, pero no es así para distinguir si es alegre o neutra, por ejemplo.

- La alegría y la neutra son emociones más rápidas que el enfado, la sorpresa y la tristeza.

7.1.8.Velocidad de locución de cada grupo fónico

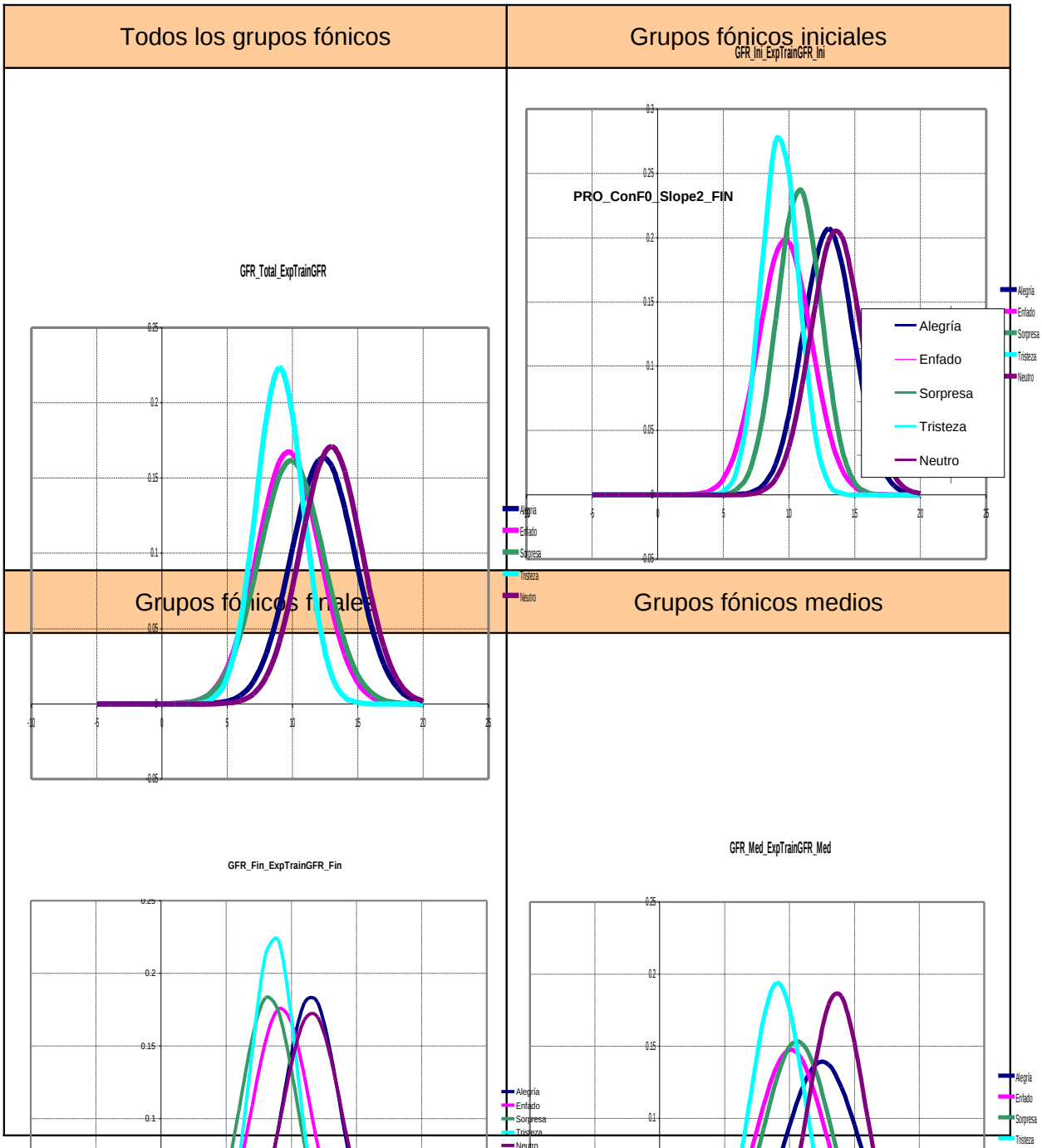


Figura 15. Modelos obtenidos para la velocidad de locución de cada grupo fónico.

La varianza de la velocidad de locución de cada grupo fónico de todas las

emociones es estrecha, pero el valor medio es muy similar, lo que hace que no podamos distinguir bien las distintas emociones. Las principales conclusiones obtenidas observando la figura, son las siguientes:

- En todos los casos (aunque cuando sólo utilizamos los grupos fónicos medios un poco menos), el modelo obtenido para la alegría y la neutra es similar.
- El modelo del enfado y la sorpresa también es similar, principalmente cuando utilizamos todos los grupos fónicos y cuando sólo utilizamos los medios.
- En general, la tristeza aparece alejada de la neutra.

7.1.9. Características apropiadas para identificar cada una de las emociones

Una vez que hemos analizado los modelos obtenidos al entrenar las distintas emociones con las distintas características, llegamos a las siguientes conclusiones:

- Sorpresa: es la emoción que presenta los modelos más diferenciados, y por tanto, permitirá una mejor identificación. Las características que mejor la identificarán son el valor medio y máximo de F0, y el rango de F0, principalmente cuando entrenamos con los grupos fónicos medios.

- Tristeza: las características que nos permitirán identificar mejor esta emoción son:

    - El valor medio de F0, principalmente cuando entrenamos con los grupos fónicos medios.
    - El valor máximo de F0 cuando entrenamos con los grupos fónicos medios o los finales.

- Alegría: los modelos obtenidos para esta emoción no están muy diferenciados de los de otras emociones. Adicionalmente, con algunas características presenta una gran varianza, que hace que se introduzca mucha distorsión. Las características que posiblemente nos permitan identificarla serán el valor máximo de F0 y su rango, utilizando los grupos fónicos medios.

- Voz neutra: la característica que nos permitirá identificar mejor esta emoción es el valor máximo de F0, principalmente cuando entrenamos con todos los grupos fónicos y sólo con los iniciales.

- Enfado: los modelos obtenidos para el enfado no están muy diferenciados de las otras emociones, por lo que es complicado poder elegir una característica que sea la que mejor lo caracterice.

Una alternativa para evaluar la bondad de estas características sería hacer un análisis basado en PCA (Principal Components Analysis) o LDA (Linear Discriminant Analysis). PCA es una técnica para simplificar un conjunto de datos, reduciendo la dimensión para el análisis. Se trata de hacer una transformación ortogonal, obteniendo un nuevo sistema de coordenadas en el que hemos reducido la dimensión de los datos. LDA es una técnica usada en estadística para encontrar la combinación lineal de los rasgos que mejor caracterizan a un cierto objeto. La combinación lineal encontrada comúnmente es utilizada en la reducción de la dimensionalidad, antes de hacer la clasificación. Las técnicas de PCA y LDA están relacionadas, pero LDA intenta modelar la diferencia entre las clases de datos, mientras que PCA no tiene en cuenta las diferencias, sino las semejanzas.

7.2.Identificación basada en estadísticos sobre el contorno de la frecuencia fundamental

En este apartado vamos a realizar una serie de experimentos de identificación en los cuales usaremos vectores de entrenamiento y clasificación con características relacionadas con el contorno de la frecuencia fundamental, para intentar descubrir qué información nos aportan sobre las distintas emociones.

7.2.1.Descripción de los experimentos

En los primeros experimentos realizados utilizaremos un vector que contenga las siguientes características de contorno de la frecuencia fundamental de cada grupo fónico:

- Valor medio de la frecuencia fundamental (F0).
- Mínimo de F0.
- Máximo de F0.
- Rango de F0.
- Pendiente de subida.
- Pendiente de bajada.

La estrategia que vamos a seguir para obtener los vectores de características es descomponiendo cada uno de los ficheros en todos sus grupos fónicos y calculando el vector asociado a cada uno de esos grupos.

A continuación se presenta la descripción de los experimentos realizados para llevar a cabo la evaluación de la contribución de la información prosódica de los grupos fónicos, según se contemplen las características de todos los grupos fónicos o sólo de los iniciales, finales o medios. La razón por la que analizamos los distintos grupos fónicos por separado se debe a que en la lengua castellana, la información prosódica se distribuye de distinta forma a lo largo de la frase.

- Experimentos A, B y C: *Basados en todos los grupos fónicos.*

En estos experimentos, los ficheros de características que emplearemos en el entrenamiento y la clasificación están formados por los vectores de todos los grupos fónicos del fichero de datos, cada uno de ellos, formado por las seis características comentadas anteriormente. Suponiendo que tenemos N grupos fónicos, el fichero de características tendría la siguiente estructura:

1	$F0_{GF1}$	Min_{GF1}	Max_{GF1}	$Rango_{GF1}$	$PendienteS_{GF1}$	$PendienteB_{GF1}$
2	$F0_{GF2}$	Min_{GF2}	Max_{GF2}	$Rango_{GF2}$	$PendienteS_{GF2}$	$PendienteB_{GF2}$
⋮						
N	$F0_{GFN}$	Min_{GFN}	Max_{GFN}	$Rango_{GFN}$	$PendienteS_{GFN}$	$PendienteB_{GFN}$

Figura 16. Estructura del fichero de características prosódicas relacionadas con F0.

- Experimentos D, E y F: *Basados en los grupos fónicos iniciales.*

Los ficheros de características están formados solamente por el vector del primer grupo fónico.

- Experimentos G, H e I: *Basados en los grupos fónicos finales.*

Los ficheros de características están formados solamente por el vector del último grupo fónico.

- Experimentos J, K y L: *Basados en los grupos fónicos medios.*

Los ficheros de características están formados por los vectores de los grupos fónicos medios, es decir, los de todos los grupos fónicos, menos el inicial y el final. Algunos de los ficheros disponibles sólo están formados por dos grupos fónicos, por lo que en estos experimentos el número de ficheros de características obtenidos será ligeramente inferior.

7.2.2.Resultados de los experimentos

En este apartado analizaremos los resultados obtenidos para los distintos experimentos explicados anteriormente. Para cada uno de ellos, mostraremos una tabla con los valores medios de los tres posibles experimentos de cada tipo, con los siguientes datos:

- Tasa de identificación de cada emoción.
- Precisión de cada emoción.
- Banda de fiabilidad de la diagonal principal.

1.1.1.34.Experimentos A, B y C:

En la siguiente tabla se muestran los resultados medios obtenidos en los *experimentos A, B* y *C* (considerando todos los grupos fónicos):

Tabla 68. Tasas de identificación para cada emoción de los experimentos A, B y C.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	33,33% (± 6,75%)	40%	6,67%	8,90%	11,10%
Enfado	13,93%	18,23% (± 5,65%)	4,60%	37,47%	25,70%
Sorpresa	2,23%		97,77% (± 2,11%)		
Tristeza	4,43%	2,23%		80% (± 5,72%)	13,33%
Neutro		2,37%	2,23%	4,77%	90,63% (± 4,27%)
PRECISIÓN	61,80%	29,02%	87,87%	61,01%	64,39%

En esta tabla observamos que:

- La emoción que mejor se identifica es la sorpresa, seguida de la voz neutra y la tristeza.
- La alegría se confunde con el enfado con una tasa de 40%. Es curioso observar que a pesar de que la alegría y el enfado en principio son emociones muy diferentes desde un punto de vista prosódico, a la hora de hacer este experimento automático obtengamos como resultado que la alegría, siendo una emoción positiva, se confunda con el enfado, que es una emoción negativa.
- El enfado se confunde principalmente con la tristeza y también con la voz neutra. La voz neutra está asociada con la seriedad, con la tristeza, así que en principio la tristeza tendrá ciertas características similares a las de la voz neutra, que serán las que hagan que el enfado se confunda con ambas.
- Hay ciertas emociones que no se confunden nunca:
    - La sorpresa nunca se confunde con el enfado, ni con la tristeza ni con la voz neutra. Y la tristeza nunca se confunde con la sorpresa. Es interesante observar que la tristeza y la sorpresa, que son las emociones que mejor se identifican, nunca se confunden la una con la otra, es decir, tienen un patrón prosódico bien definido y unívoco.
    - La voz neutra nunca se confunde con la alegría.
- La sorpresa tiene la precisión más elevada y también es la que tiene la mayor tasa de identificación, por lo que podemos concluir que es la emoción que mejor se reconoce empleando características prosódicas relacionadas con F0.
- La alegría, la tristeza y la voz neutra tienen una precisión similar. La alegría se confunde con el enfado y sólo se identifica en un 33,33% de los casos, sin embargo, fijándonos en la precisión, vemos que ésta es mayor que la de la tristeza, que se identifica en un 80% de los casos. Es decir, la alegría no se identifica en muchas ocasiones, pero la probabilidad de qué cuando una emoción se reconozca como alegría, verdaderamente lo sea, es elevada.

1.1.1.35.Experimentos D, E y F:

Los resultados obtenidos en los *experimentos D, E* y *F* (considerando sólo los grupos fónicos iniciales), son los que se muestran a continuación:

Tabla 69. Tasas de identificación para cada emoción de los *experimentos D, E* y *F*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	33,33% (± 6,75%)	44,43%	13,30%	4,47%	4,47%
Enfado	9,53%	46,80% (± 7,3%)		34,43%	9,20%
Sorpresa	6,67%		93,33% (± 3,57%)		
Tristeza	4,43%	31,10%		53,30% (± 7,14%)	11,10%
Neutro		16,50%	2,23%	9,37%	71,90% (± 6,58%)
PRECISIÓN	61,77%	33,71%	85,73%	52,48%	74,38%

Con los resultados obtenidos en la tabla podemos afirmar que:

- La mayor tasa de identificación la tiene la sorpresa, seguido de la voz neutra.
- En este caso disminuyen las tasas de identificación de la sorpresa, la tristeza y la voz neutra (respecto al experimento anterior), lo que en principio parece lógico, ya que estamos trabajando con muchos menos datos de entrenamiento. La tasa que más se reduce es la de la tristeza (de un 80% a un 53,3%). Esto se puede deber a que la información prosódica relacionada con F0 contenida en la parte inicial de las frases tristes, no sea una característica relevante de esta emoción.
- Sin embargo, la tasa de identificación del enfado aumenta considerablemente y deja de confundirse con la neutra (la tasa de confusión disminuye de un 25,7% a un 9,2%). Esto se puede deber a que esta emoción posea mucha información prosódica en la parte inicial de la frase o que la información de la parte inicial difiera bastante de la de la neutra.
- La alegría se confunde con el enfado un 44,43% y se identifica un 33,33%. Estas tasas de identificación y confusión nos hacen pensar que las características prosódicas de la alegría y el enfado, son similares, a pesar de que una sea positiva y la otra negativa.
- Fijándonos en la alegría y la sorpresa, respecto al experimento anterior, observamos que sube tanto la tasa de confusión de la alegría con la sorpresa (de 6,67% a 13,3%) y la de la sorpresa con la alegría (de 2,23% a 6,67%). Esto podemos atribuirlo a que la información prosódica de la alegría y la sorpresa se parezca más en la parte inicial de una frase.
- El enfado se identifica un 46,8% y se confunde con la tristeza un 34,43%.
- Al igual que en el caso en el que utilizamos la información de todos los grupos fónicos, la sorpresa nunca se confunde con el enfado, ni la tristeza, ni la voz neutra, la tristeza nunca se confunde con la sorpresa y la neutra nunca se confunde con la alegría. Adicionalmente, en este caso, el enfado nunca se confunde con la sorpresa.
- La sorpresa tiene la precisión más elevada y también es la que tiene mayor tasa de identificación, por lo que podemos decir que es la emoción que mejor se

reconoce cuando utilizamos sólo los grupos fónicos iniciales.

- La precisión de la alegría es elevada, al igual que en los experimentos anteriores, a pesar de que se confunde con el enfado.
- Aunque la tasa de identificación del enfado hemos visto que mejoraba en este caso, la precisión es menor, ya que se confunde mucho con la alegría y con la tristeza.

1.1.1.36.Experimentos G, H e I:

En la siguiente tabla se muestran los resultados medios obtenidos en los *experimentos G, H* e *I* (considerando sólo los grupos fónicos finales):

Tabla 70. Tasas de identificación para cada emoción de los *experimentos G, H* e *I*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	20,03% (± 5,73%)	13,37%	15,53%	33,33%	17,77%
Enfado	23,67%	11,27% (± 4,63%)	11,43%	16,33%	37,27%
Sorpresa	15,57%		84,43% (± 5,19%)		
Tristeza		2,23%		80% (± 5,72%)	17,77%
Neutro	4,60%	7,00%		46,37%	42,03% (± 7,23%)
PRECISIÓN	31,37%	33,27%	75,79%	45,45%	36,60%

En la tabla podemos observar que:

- La mayor tasa de identificación es para la sorpresa, que siempre se identifica, salvo en un 15,57% que se confunde con la alegría.
- Junto con la sorpresa, está la tristeza, que se identifica con la misma tasa (80%) que cuando utilizábamos las características de todos los grupos fónicos.
- La tasa de identificación de la voz neutra disminuye bastante respecto a los dos casos anteriores y se confunde con la tristeza (en un 46,37%), por lo que es posible que la información prosódica de los grupos fónicos finales de la neutra y la tristeza sea similar, algo que en principio puede parecer razonable, ya que la voz neutra se suele asociar con la tristeza.

- Tanto la alegría como el enfado no se identifican, pero se confunden con todas las demás emociones. La alegría principalmente se confunde con la tristeza y el enfado con la voz neutra.
- Al igual que en los experimentos anteriores (*A-F)* la sorpresa nunca se confunde con el enfado, ni la tristeza, ni la voz neutra; y la tristeza nunca se confunde con la alegría ni con la sorpresa. En cambio, en este caso la neutra nunca se confunde con la sorpresa y sí con la alegría (aunque las tasas de confusión con la sorpresa en los otros experimentos, así como el de confusión con la alegría en éste, son muy bajos).
- La sorpresa tiene la precisión más elevada y también es la que tiene la mayor tasa de identificación, por lo que podemos concluir que también es la emoción que mejor se reconoce en este caso.
- El resto de emociones tienen una baja precisión.

1.1.1.37.Experimentos J, K y L:

Los resultados obtenidos en los *experimentos J, K* y *L* (considerando sólo los grupos fónicos medios) son los que se muestran en la tabla que aparece a continuación:

Tabla 71. Tasas de identificación para cada emoción de los *experimentos J, K* y *L*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	51,27% (± 7,68%)	10,27%	7,70%	10,27%	20,53%
Enfado	20,97%	2,57% (± 2,46%)	7,70%	37,83%	31,00%
Sorpresa	5,13%		94,87% (± 3,39%)		
Tristeza		2,57%		92,30% (± 4,1%)	5,13%
Neutro	2,57%			2,57%	94,87% (± 3,48%)
PRECISIÓN	64,14%	16,67%	86,03%	64,56%	62,60%

En esta tabla observamos que:

- Utilizando sólo la información de los grupos fónicos medios aumenta la tasa de identificación de la alegría, la tristeza y la voz neutra. Sin embargo, la de la

sorpresa disminuye, pero no significativamente.

- Este es el único caso (de los que llevamos analizados hasta ahora de prosodia) en los que la alegría se identifica (51,27%) y no se confunde con otra emoción.
- El enfado, sin embargo, sigue confundiéndose con otras emociones, principalmente con la tristeza y la voz neutra.
- Al igual que en los experimentos anteriores, la sorpresa nunca se confunde con el enfado, ni con la tristeza, ni con la neutra; y la tristeza nunca se confunde con la sorpresa, ni tampoco (sólo en estos experimentos) con la alegría.
- La neutra nunca se confunde ni con el enfado ni con la sorpresa.
- Aún siendo la tasa de identificación igual para la sorpresa y la neutra (94,87%), fijándonos en la precisión, observamos que la sorpresa tiene una precisión mayor (86,03% frente a 62,6%), por lo que sigue siendo la emoción que mejor se reconoce.
- La precisión de la alegría, la tristeza y la neutra es similar, un poco mayor de 60%; y la del enfado es baja.

7.2.3.Conclusiones de los experimentos

- La alegría sólo se identifica en el caso en el que utilizamos los grupos fónicos medios (51,27%).
- El enfado sólo se identifica en el caso en el que utilizamos los grupos fónicos iniciales (46,8%), pero en este caso, la alegría se confunde con el enfado (con un 44,43%), lo que hace que la precisión sea baja. El hecho de que la alegría se confunda con el enfado se puede deber a que la información prosódica de estas emociones en la parte inicial de la frase sea similar, a pesar de que se trate de una emoción positiva y otra negativa.
- La sorpresa es la emoción que mayor tasa de identificación tiene en todos los casos. La mayor tasa de identificación y la mayor precisión, se obtienen cuando utilizamos todos los grupos fónicos (97,77% y 87,87%, respectivamente). La menor tasa de identificación y menor precisión, se obtienen cuando sólo usamos los grupos fónicos finales (84,43% y 75,79%, respectivamente). Con estos datos

podemos suponer que la información prosódica relacionada con F0 de la sorpresa se encuentra principalmente en los grupos fónicos iniciales y medios.

- La tristeza tiene la mayor tasa de identificación cuando usamos sólo los grupos fónicos medios (92,3%), mientras que la menor se da cuando sólo se usan los grupos fónicos iniciales (53,33%). Respecto a su precisión, la mayor se obtiene con los grupos fónicos medios (64,56%) y la menor con los grupos fónicos finales (45,45%). Podemos concluir por tanto, que en los grupos fónicos medios es donde posiblemente se encuentra la información prosódica relacionada con F0 de la tristeza.

- La voz neutra tiene la mayor tasa de identificación cuando se utilizan los grupos fónicos medios (94,87%) y la menor, con los grupos fónicos finales (42,03%), confundiéndose en este caso con la tristeza (con un 46,37%), por lo que podemos suponer que la información prosódica relacionada con F0 de la parte final de la frase de la neutra y la tristeza son similares. Su mayor precisión se obtiene cuando utilizamos los grupos fónicos iniciales (74,38%).

- En todos los experimentos anteriores (con todos los grupos fónicos, sólo con los iniciales, sólo con los finales y sólo con los medios) en los que utilizamos un vector con seis características, se cumple que:

  - La sorpresa nunca se confunde con el enfado, ni con la tristeza, ni con la neutra.
  - La tristeza nunca se confunde con la sorpresa y se confunde muy poco con la alegría, como cabía esperar, dadas las diferencias prosódicas existentes entre estas emociones.
  - La neutra casi no se confunde con la alegría, ni con la sorpresa.
  - De los datos anteriores podemos concluir que la sorpresa y la tristeza posiblemente tengan características prosódicas relacionadas con F0 diferenciadas una respecto a la otra, que hacen que no se confundan entre ellas. Y lo mismo podemos decir sobre la sorpresa y la voz neutra. Es decir, obtenemos un patrón unívoco para la sorpresa, que nos permite identificarla con una tasa elevada y evitar que se confunda con la tristeza y la voz neutra.

7.3.Relevancia del valor medio de la frecuencia fundamental

Con el objetivo de determinar la relevancia del valor medio de F0 a la hora de identificar emociones, vamos a realizar una serie de experimentos en los que prescindamos de dicho valor, que son los que se describen a continuación.

7.3.1.Descripción de los experimentos

Estos experimentos son similares a los realizados en el apartado anterior (*apartado 1.1.1.33*), salvo que en este caso se prescinde del valor medio de F0 en el vector de características. La razón por la que prescindimos de esta característica es que se supone que se trata de una de las características cuya contribución es mayor a la hora de identificar las emociones, y de esta forma podremos comprobar si lo es o no, analizando y comparando, las tasas de identificación considerándolo o no en el vector de características.

Los experimentos realizados en este apartado son los siguientes, que se diferencian en los grupos fónicos que consideremos en el vector de características:

- Experimentos M, N y O: *Basados en todos los grupos fónicos.*
- Experimentos P, Q y R: *Basados en los grupos fónicos iniciales.*
- Experimentos S, T y U: *Basados en los grupos fónicos finales.*
- Experimentos V, W y X: *Basados en los grupos fónicos medios.*

7.3.2.Resultados de los experimentos

En este apartado analizaremos los resultados obtenidos para los experimentos en los que prescindimos del valor medio de F0 en el vector de características. En cada uno de los ellos compararemos los resultados obtenidos con los que obtuvimos con el mismo tipo de experimento, pero considerando el valor medio de F0.

Para cada uno de los experimentos, mostraremos una tabla que contenga la tasa

de identificación media de cada emoción, la precisión de cada emoción y la banda de fiabilidad de la diagonal principal.

Además, para poder analizar las diferencias en la precisión considerando o no el valor medio de F0, se mostrará en una tabla la precisión para cada emoción en ambos casos.

1.1.1.38.Experimentos M, N y O

En la siguiente tabla se muestran los resultados medios obtenidos en los *experimentos M, N* y *O* (considerando todos los grupos fónicos):

Tabla 72. Tasas de identificación para cada emoción de los *experimentos M, N* y *O*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	26,67% (± 6,33%)	35,57%	13,30%	17,80%	6,70%
Enfado	9,17%	18,23% (± 5,65%)	7,00%	49,20%	16,33%
Sorpresa	2,23%		97,77% (± 2,11%)		
Tristeza	4,43%	2,23%		73,33% (± 6,33%)	19,97%
Neutro		4,60%	2,23%	21,43%	71,77% (± 6,59%)
PRECISIÓN	62,75%	30,07%	81,27%	45,33%	62,53%

En esta tabla observamos que:

- La emoción que tiene una mayor tasa de identificación es la sorpresa, con una tasa igual (97,77%) a la que obteníamos cuando considerábamos la media de F0 en el vector de características.
- Las siguientes emociones que tienen mayor tasa de identificación son la tristeza y la voz neutra.
- La alegría se identifica sólo un 26,67% y se confunde con el enfado en un 35,57%.
- El enfado se confunde con la tristeza (49,2%).
- Hay ciertas emociones que nunca se confunden:
    - La sorpresa nunca se confunde con el enfado, ni con la tristeza, ni con la

neutra.

- La tristeza nunca se confunde con la sorpresa.
- La neutra nunca se confunde con la alegría.

- La sorpresa tiene la precisión más elevada y también es la que tiene la mayor tasa de identificación, por lo que se trata de la emoción que mejor se reconoce.
- La alegría y la voz neutra tienen precisiones similares, a pesar de que la tasa de identificación de la neutra (71,77%) es mucho mayor que la de la alegría (26,67%).
- La tristeza, a pesar de tener una elevada tasa de identificación (73,33%), tiene una precisión baja (45,33%), debido a que el enfado se confunde con ella.

Si comparamos los resultados obtenidos en estos experimentos (*M, N* y *O*) y en los que considerábamos el valor medio de F0 en el vector de características para todos los grupos fónicos (*A, B* y *C*), podemos llegar a las siguientes conclusiones:

- La alegría se identifica mejor cuando utilizamos el valor medio de F0. Aunque si no utilizamos ese dato, se confunde menos con el enfado (un 40% frente a un 35,57%) y la confusión se reparte un poco entre todas las emociones.
- La confusión del enfado con la tristeza es mayor si no utilizamos el valor medio de F0 (37,47% frente a un 49,2%).
- La tasa de identificación de la sorpresa es la misma utilicemos o no el valor medio de F0 (97,77%).
- La tasa de identificación de la tristeza y la neutra se ven reducidos al no considerar el valor medio de F0 (la tristeza de un 80% a un 73,33% y la neutra de un 90,63% a un 71,77%). La confusión de la tristeza con la neutra aumenta al no considerarlo (de un 13,33% a un 19,97%) y la de la neutra con la tristeza también (de un 4,77% a un 21,43%).
- La precisión de cada una de las emociones no sufre grandes variaciones. La precisión de la alegría y el enfado aumentan ligeramente cuando no utilizamos el valor medio de F0. En cambio, en la sorpresa, la tristeza y la neutra la precisión es mayor cuando sí que utilizamos el valor medio de F0. La siguiente tabla muestra los valores para ambos casos, considerando o no el valor medio de F0.

Tabla 73. Precisión de cada emoción de los experimentos en los que consideramos o no el valor medio de F0 (todos los grupos fónicos).

	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Con_MedF0	61,80%	29,02%	87,87%	61,01%	64,39%
Sin_MedF0	62,75%	30,07%	81,27%	45,33%	62,53%

1.1.1.39. Experimentos P, Q y R

Podemos evaluar los resultados medios obtenidos en los *experimentos P, Q* y *R* (considerando sólo los grupos fónicos iniciales) en la tabla que aparece a continuación:

Tabla 74. Tasas de identificación para cada emoción de los *experimentos P, Q* y *R*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	22,20% (± 5,95%)	51,13%	11,10%	11,10%	4,47%
Enfado	2,37%	44,43% (± 7,27%)	2,37%	34,43%	16,33%
Sorpresa	6,67%		91,13% (± 4,07%)	2,23%	
Tristeza	4,43%	20,00%		46,67% (± 7,14%)	28,90%
Neutro		16,50%	2,23%	16,53%	64,77% (± 6,49%)
PRECISIÓN	62,24%	33,64%	85,30%	42,05%	56,58%

Analizando los datos mostrados en la tabla podemos afirmar que:

- La emoción con mayor tasa de identificación es la sorpresa.
- A continuación están la neutra, la tristeza y el enfado, pero con tasas bastante más bajas que la de la sorpresa.
- Respecto al anterior experimento, la tasa de identificación del enfado aumenta considerablemente (de un 18,23% a un 44,43%), por lo que es probable, que la información prosódica del enfado se encuentre en la parte inicial de la frase.
- Al igual que sucedía cuando considerábamos el valor medio de la F0, la tasa de identificación de la tristeza disminuye mucho cuando utilizamos sólo los grupos fónicos iniciales en comparación con cuando utilizamos todos los grupos fónicos (de un 73,33% a un 46,67%), por lo que es probable que la información prosódica

de la tristeza no se encuentre en la parte inicial de la frase.

- En este caso disminuyen las tasas de identificación de la sorpresa, la tristeza y la voz neutra (respecto al experimento anterior), lo que es puede ser lógico, ya que estamos trabajando con muchos menos datos de entrenamiento.
- La confusión de la alegría con el enfado aumenta respecto al experimento anterior (de un 35,57% a un 51,13%).
- Al igual que en el experimento anterior, la sorpresa nunca se confunde con el enfado ni con la neutra; la tristeza nunca se confunde con la sorpresa; y la neutra nunca se confunde con la alegría.
- La sorpresa tiene la precisión más elevada y también es la que tiene la mayor tasa de identificación, por lo que podemos decir que es la emoción que mejor se reconoce.
- Fijándonos en la precisión, la segunda emoción que tiene una mayor precisión es la alegría, a pesar de que se confunde con el enfado con una tasa de un 51,13%. Esto quiere decir que cuando una emoción se identifica como alegría, es bastante probable que verdaderamente lo sea.

Si comparamos los resultados obtenidos en estos experimentos (*P, Q* y *R*) y en los que considerábamos el valor medio de F0 en el vector de características para los grupos fónicos iniciales (*D, E* y *F*), podemos llegar a las siguientes conclusiones:

- La confusión de la alegría con el enfado es mayor si no utilizamos el valor medio de F0 (la tasa aumenta de un 44,43% a un 51,13%).
- La tasa de identificación del enfado, la sorpresa, la tristeza y la neutra es mayor cuando utilizamos el valor medio de F0, pero no es un aumento significativo:
    - *Enfado:* la tasa de identificación disminuye de un 46,8% a un 44,43%.
    - *Sorpresa:* la tasa de identificación disminuye de un 93,33% a un 91,13%.
    - *Tristeza:* la tasa de identificación disminuye de un 53,3% a un 46,67%.
    - *Neutra:* la tasa de identificación disminuye de un 71,9% a un 64,77%.
- La precisión no sufre grandes cambios dependiendo si consideramos o no el valor medio de F0. En todos los casos, salvo para la alegría, es mayor al considerar el

valor medio de F0, pero los mayores incrementos se producen para la tristeza y, principalmente, para la neutra, como podemos ver en la siguiente tabla:

Tabla 75 Precisión de cada emoción de los experimentos en los que consideramos o no el valor medio de F0 (grupos fónicos iniciales).

	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Con_MedF0	61,77%	33,71%	85,73%	52,48%	74,38%
Sin_MedF0	62,24%	33,64%	85,30%	42,05%	56,58%

1.1.1.40.Experimentos S, T y U

En la siguiente tabla se muestran los resultados medios obtenidos en los *experimentos S, T* y *U* (considerando sólo los grupos fónicos finales):

Tabla 76. Tasas de identificación para cada emoción de los *experimentos S, T* y *U*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	24,43% (± 6,15%)	11,13%	13,30%	31,10%	20,03%
Enfado	23,67%	11,40% (± 4,65%)	11,43%	11,57%	41,90%
Sorpresa	13,33%		86,67% (± 4,86%)		
Tristeza		2,23%		80% (± 5,72%)	17,77%
Neutro	4,60%	4,60%		46,53%	44,27% (± 7,27%)
PRECISIÓN	37,00%	38,82%	77,80%	47,28%	35,71%

En esta tabla observamos que:

- Las emociones que tienen una mayor tasa de identificación son la sorpresa y la tristeza.
- La neutra tiene una tasa de identificación similar a la de confusión con la tristeza (44,27% frente a 46,53%).
- El enfado se confunde con la neutra (41,9%).
- La alegría se confunde principalmente con la tristeza (31,10%), pero también se identifica con una tasa similar a la que se confunde con la neutra (24,43% de identificación, frente a 20,03% de confusión con la neutra).

- Al igual que en los casos anteriores, la sorpresa nunca se confunde con el enfado, ni con la tristeza, ni con la neutra; la tristeza nunca se confunde con la alegría, ni con la sorpresa; y la neutra, en este caso, nunca se confunde con la sorpresa.
- La precisión más elevada (y con mucha diferencia) se obtiene para la sorpresa.
- La precisión de la tristeza es bastante inferior a la de la sorpresa, a pesar de que la tasa de identificación era similar, debido a que la neutra y la alegría se confunden con ella.
- La alegría, el enfado y la voz neutra tienen una precisión similar.

Si comparamos los resultados obtenidos en estos experimentos (*S, T* y *U*) y en los que utilizábamos el valor medio de F0 en el vector de características para los grupos fónicos finales (*G, H* e *I*), podemos llegar a las siguientes conclusiones:

- La alegría se confunde menos con la tristeza y más con la neutra cuando no consideramos el valor medio de F0 (la confusión con la tristeza se reduce de un 33,33% a un 31,1% y la confusión con la neutra aumenta de un 17,7% a un 20,03%).
- El enfado, al igual que la alegría, se confunde menos con la tristeza y más con la neutra cuando no consideramos el valor medio de F0 (la confusión con la tristeza se reduce de un 16,33% a un 11,57% y la confusión con la neutra aumenta de un 37,27% a un 41,9%).
- La sorpresa y la neutra se identifican mejor cuando no consideramos el valor medio de F0, pero el aumento no es significativo (la tasa de identificación de la sorpresa aumenta de un 84,43% a un 86,67% y la de la neutra de un 42,03% a un 44,27%).
- La tasa de identificación de la tristeza es la misma consideremos o no el valor medio de F0 (80%).
- La precisión de todas las emociones (salvo de la voz neutra) aumenta al no considerar el valor medio de F0, como podemos observar en la siguiente tabla:

Tabla 77. Precisión de cada emoción de los experimentos en los que consideramos o no el valor medio de F0 (grupos fónicos finales).

	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Con_MedF0	31,37%	33,27%	75,79%	45,45%	36,60%
Sin_MedF0	37,00%	38,82%	77,80%	47,28%	35,71%

1.1.1.41.Experimentos V, W y X

En la tabla que aparece a continuación podemos observar los resultados medios obtenidos en los *experimentos V, W* y *X* (considerando sólo los grupos fónicos medios):

Tabla 78. Tasas de identificación para cada emoción de los *experimentos V, W* y *X*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	51,27% (± 7,68%)	10,27%	7,70%	5,13%	25,67%
Enfado	23,53%		7,70%	32,70%	36,10%
Sorpresa	2,57%	2,57%	94,87% (± 3,39%)		
Tristeza		5,13%		89,73% (± 4,67%)	5,13%
Neutro	2,57%	2,77%		5,53%	89,10% (± 4,92%)
PRECISIÓN	64,14%		86,03%	67,42%	57,12%

En esta tabla observamos que:

- La mayor tasa de identificación la tienen la sorpresa, la tristeza y la voz neutra.
- El enfado nunca se identifica.
- Las emociones que nunca se confunden en este caso son:
    - La sorpresa nunca se confunde con la tristeza, ni con la neutra.
    - La tristeza nunca se confunde con la alegría, ni con la sorpresa.
    - La neutra nunca se confunde con la sorpresa.
- La sorpresa tiene la precisión más elevada y la mayor tasa de identificación, por lo que es la emoción que mejor se reconoce.

- La alegría tiene una precisión similar a la tristeza y superior a la neutra, a pesar que su tasa de identificación es bastante menor.

Si comparamos los resultados obtenidos en estos experimentos (*V, W* y *X*) y en los que consideramos el valor medio de F0 en el vector de características para los grupos fónicos medios (*J, K* y *L*), podemos llegar a las siguientes conclusiones:

- La tasa de identificación de la alegría y de la sorpresa permanecen invariables consideremos o no el valor medio de F0.
- La tasa de identificación de la tristeza y la neutra se ven reducidas, aunque no significativamente, cuando no consideramos el valor medio de F0 (en la tristeza disminuye de un 92,3% a un 89,73% y en la neutra de un 94,87% a un 89,1%).
- El enfado se confunde menos con la tristeza y más con la neutra cuando no consideramos el valor medio de F0 (la confusión con la tristeza se reduce de un 37,83% a un 32,7% y la confusión con la neutra aumenta de un 31% a un 36,1%).
- A pesar de que la tasa de identificación de la tristeza se ve reducida al no considerar el valor medio de F0, la precisión aumenta (de un 64,56% a un 67,42%).
- La precisión de la alegría y la sorpresa permanecen constantes consideremos o no el valor medio de F0 (al igual que pasaba con la tasa de identificación).
- La precisión del enfado y la neutra se ven reducidas al no utilizar el valor medio de F0 (el enfado no se identifica nunca, mientras que la neutra disminuye de un 62,6% a un 57,12%).

7.3.3.Conclusiones de los experimentos

- El enfado se identifica mejor cuando utilizamos los grupos fónicos iniciales (51,13%), por lo que puede ser, que la información prosódica del enfado se encuentre al principio de la frase.
- Cuando usamos los grupos fónicos iniciales aumenta la tasa de confusión de la alegría con el enfado (51,13%). Esto se puede deber a que la información prosódica al inicio de la frase es similar para la alegría y el enfado.
- La alegría sólo se identifica en el caso en el que utilizamos los grupos fónicos

medios (con un 51,27%).

- La sorpresa es la emoción que tiene la mayor tasa de identificación en todos los casos. En el experimento en el que utilizamos todos los grupos fónicos es en el que se obtiene la mayor tasa (97,77%). Sin embargo, la mayor precisión se obtiene para el experimento en el que sólo utilizamos los grupos fónicos medios (86,03%). La menor tasa de identificación de la sorpresa se obtiene cuando solamente utilizamos los grupos fónicos finales (86,67%). Para este caso también es para el que se obtiene la menor precisión (77,8%). Esto nos hace pensar que en la parte final de la frase es donde existe menos información prosódica relacionada con F0 que caracterice a la sorpresa.

- La tristeza tiene una mayor tasa de identificación para el experimento en el que solamente utilizamos los grupos fónicos medios (89,73%), siendo ese experimento también en el que se obtiene una mayor precisión (67,42%). Su menor tasa de identificación se obtiene cuando solamente usamos los grupos fónicos iniciales (46,67%) y también, en este caso, obtenemos la menor precisión (42,05%). Con estos datos podemos intuir que la información prosódica relacionada con F0 de la tristeza se encuentra principalmente en la parte central de la frase.

- La voz neutra, al igual que la tristeza, tiene la mayor tasa de identificación para el experimento en el que solamente utilizamos los grupos fónicos medios (89,1%). Sin embargo, la mayor precisión se consigue cuando utilizamos todos los grupos fónicos (62,53%). La menor tasa de identificación de la neutra se obtiene cuando solamente usamos los grupos fónicos finales (44,27%), confundiéndose en este caso con la tristeza (46,53%).

- En todos los experimentos anteriores (con todos los grupos fónicos, sólo con los iniciales, sólo con los finales y sólo con los medios) en los que utilizamos un vector con cinco características, se cumple que:

    - La sorpresa nunca se confunde con el enfado, ni con la tristeza, ni con la neutra (utilizando sólo los grupos fónicos iniciales, la sorpresa se confunde en un 2,23% con la tristeza y en el caso de utilizar los grupos fónicos medios, la sorpresa se confunde con el enfado en un 2,57%).

- La tristeza nunca se confunde con la sorpresa y en pocas ocasiones con la alegría.
- La neutra prácticamente no se confunde con la alegría, ni con la sorpresa.
- De los datos anteriores podemos concluir que la sorpresa y la tristeza tienen características prosódicas diferenciadas entre ellas, que hacen que no se confundan mutuamente. Y lo mismo podemos decir sobre la sorpresa y la voz neutra. Por tanto, el patrón obtenido para la sorpresa nos permite identificar perfectamente esta emoción frente a la tristeza y a la voz neutra.

- Del análisis realizado considerando o no el valor medio de F0 en el vector de características, podemos obtener las siguientes conclusiones:
  - En general, los resultados obtenidos son mejores cuando consideramos el valor medio de F0 (las tasas de identificación aumentan y las tasas de confusión disminuyen), por lo que esto implica que si que se trata de una característica que tenga una importante relevancia a la hora de reconocer las emociones.
  - El único caso en el que obtenemos mejores resultados al no considerar el valor medio de F0 es trabajando sólo con los grupos fónicos finales. En este caso, la precisión de todas las emociones es mayor sin el valor medio de F0 en el vector de características.

7.4.Experimentos con la velocidad de locución de la frase

Una vez analizados los resultados obtenidos en los experimentos realizados con características relacionadas con el contorno de F0, vamos a realizar y analizar una serie de experimentos con características relacionadas con el ritmo, para evaluar como influye éste en la identificación de emociones. Concretamente, en este apartado vamos a estudiar la influencia de la velocidad de locución de la frase.

7.4.1.Descripción de los experimentos

La característica considerada en los primeros experimentos realizados con características relacionadas con el ritmo, es la velocidad de locución de toda la frase, por lo que no podremos hacer distinción en grupos fónicos, como hemos hecho en los experimentos realizados anteriormente. Por tanto, el único tipo de experimentos es el que se describe a continuación:

- Experimentos Y, Z y AA:

  En estos experimentos, los ficheros de características que emplearemos en el entrenamiento y la clasificación están formados por un único vector, formado a su vez por un único valor, que es la velocidad de locución de la frase en cuestión.

7.4.2.Resultados de los experimentos

Los resultados obtenidos para estos experimentos, en los que sólo consideramos la velocidad de locución de toda la frase son los que se muestran en la siguiente tabla, en la que se incluyen la tasa de identificación media de cada una de las emociones, así como su precisión y las bandas de fiabilidad.

Tabla 79. Tasas de identificación para cada emoción de los *experimentos Y, Z* y *AA*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	26,67% (± 6,33%)	2,23%	11,13%	2,23%	57,77%
Enfado		2,37% (± 2,23%)	42,23%	41,77%	13,67%
Sorpresa	6,67%	8,90%	22,23% (± 5,95%)	57,80%	4,47%
Tristeza		2,23%	17,80%	80% (± 5,72%)	
Neutro	26,20%		4,43%		69,37% (± 6,75%)
PRECISIÓN	44,79%	15,04%	22,73%	44,00%	47,75%

Observando los resultados que aparecen en la tabla, podemos afirmar que:

- La emoción que mayor tasa de identificación tiene es la tristeza, seguida de la voz neutra.
- La alegría, el enfado y la sorpresa se confunden con otras emociones:
    - La alegría se confunde con la neutra (57,77%).
    - El enfado se confunde con la sorpresa (42,23%) y con la tristeza (41,77%).
    - La sorpresa se confunde con la tristeza (57,8%).
- El enfado nunca se confunde con la alegría, por lo que se intuye que el ritmo de la locución de estas emociones es muy diferente.
- La tristeza nunca se confunde con la alegría, ni con la neutra; y la neutra nunca se confunde con el enfado ni con la tristeza. Con estos datos concluimos que el ritmo de la locución de la tristeza y la neutra difieren bastante, lo que hace que nunca se confundan entre ellas.
- En general, las precisiones obtenidas en estos experimentos son bajas. Así, a pesar de que la tasa de identificación de la tristeza es elevada (80%), observamos que la precisión no lo es (44%) debido a que tanto el enfado como la sorpresa se confunden con la tristeza.
- La alegría se confunde con la neutra, pero, a pesar de ello, la precisión de la alegría es similar (incluso mayor) a la de la tristeza.

7.4.3.Conclusiones de los experimentos

Utilizando la velocidad de locución de la frase solamente se identifican la tristeza y la voz neutra (obteniéndose un patrón unívoco, ya que nunca se confunden la una con la otra), pero las precisiones son bajas, ya que el enfado y la sorpresa se confunden con la tristeza, y la alegría se confunde con la neutra.

7.5.Experimentos con la velocidad de cada grupo fónico

Mediante los experimentos realizados en el apartado anterior hemos analizado la influencia de la velocidad de locución de la frase en la identificación de emociones. En este apartado estudiaremos también la influencia de una característica relacionada con el ritmo, pero en este caso se trata de la velocidad de cada uno de los grupos fónicos en los que podemos dividir la frase.

7.5.1.Descripción de los experimentos

A continuación se presenta la descripción de los experimentos que nos permitirán la evaluación de la contribución de la información prosódica relacionada con el ritmo de cada grupo fónico, según se contemplen las características de todos los grupos fónicos o sólo de los iniciales, finales o medios. Para ello, descompondremos cada uno de los ficheros en todos sus grupos fónicos y calcularemos la velocidad asociada a cada uno de ellos. Por tanto, los experimentos realizados en este apartado son los siguientes:

- Experimentos AB, AC y AD: *Basados en todos los grupos fónicos.*

  En estos experimentos, los ficheros de características que emplearemos en el entrenamiento y la clasificación están formados por los vectores de todos los grupos fónicos del fichero de datos, cada uno de ellos, formado por una única característica, que es la velocidad del grupo fónico en cuestión.

- Experimentos AE, AF y AG: *Basados en los grupos fónicos iniciales.*

  Los ficheros de características están formados solamente por el vector del primer grupo fónico, es decir, por la velocidad de locución del grupo fónico inicial.

- Experimentos AH, AI y AJ: *Basados en los grupos fónicos finales.*

  Los ficheros de características están formados solamente por el vector del último grupo fónico, es decir, por la velocidad de locución del grupo fónico final.

- Experimentos AK, AL y AM: *Basados en los grupos fónicos medios.*

Los ficheros de características están formados por los vectores de los grupos fónicos medios.

7.5.2.Resultados de los experimentos

En este apartado analizaremos los resultados obtenidos para los distintos experimentos explicados anteriormente. Para cada uno de ellos, mostraremos una tabla con los siguientes datos:

- Tasa de identificación media de cada emoción.
- Precisión de cada emoción.
- Banda de fiabilidad de la diagonal principal.

1.1.1.42.Experimentos AB, AC y AD:

Los resultados de los *experimentos AB, AC* y *AD* (considerando todos los grupos fónicos) son los que se muestran en la siguiente tabla:

Tabla 80. Tasas de identificación para cada emoción de los *experimentos AB, AC* y *AD*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	22,23% (± 5,95%)	6,67%	8,90%	6,67%	55,57%
Enfado	4,77%	16,03% (± 5,37%)	21,40%	48,90%	8,90%
Sorpresa		24,47%	24,43% (± 6,15%)	35,53%	15,53%
Tristeza		6,67%	11,13%	82,20% (± 5,47%)	
Neutro	20,60%	2,37%	2,23%		74,77% (± 6,36%)
PRECISIÓN	46,71%	28,53%	35,88%	47,43%	48,31%

Con los datos mostrados en la tabla podemos afirmar que:

- La mayor tasa de identificación se obtiene para la tristeza, seguida de la voz neutra.
- La alegría se confunde con la neutra (55,57%), lo que nos puede hacer pensar que

la velocidad de locución de los distintos grupos fónicos es similar en ambas emociones.

- El enfado y la sorpresa se confunden con la tristeza (con un 48,9% y un 35,53%, respectivamente).
- Algunas emociones nunca se confunden, lo que se puede deber a que las velocidades de sus grupos fónicos sean muy diferentes:
    - La sorpresa nunca se confunde con la alegría.
    - La tristeza nunca se confunde con la alegría, ni con la neutra.
    - La neutra nunca se confunde con la tristeza.
- A pesar de que las tasas de identificación para la tristeza y la neutra son elevadas (82,2% y 74,77%, respectivamente), observamos que la precisión no es muy elevada, debido a que otras emociones se confunden con ellas: la alegría se confunde con la neutra, y el enfado y la sorpresa se confunden con la tristeza.

1.1.1.43.Experimentos AE, AF y AG:

En la siguiente tabla se muestran los resultados medios obtenidos en los *experimentos AE, AF* y *AG* (considerando sólo los grupos fónicos iniciales):

Tabla 81. Tasas de identificación para cada emoción de los *experimentos AE, AF* y *AG*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	33,33% (± 6,75%)		26,67%	15,53%	24,47%
Enfado	4,43%	6,97% (± 3,73%)	27,60%	60,93%	
Sorpresa	2,23%	4,47%	22,23% (± 5,95%)	60%	11,10%
Tristeza		6,67%	6,67%	86,63% (± 4,87%)	
Neutro	54,63%		16,33%	6,80%	22,23% (± 6,09%)
PRECISIÓN	35,22%	38,49%	22,35%	37,68%	38,47%

En esta tabla observamos que:

- La única emoción que se identifica es la tristeza, con un 86,63%.
- El resto de emociones no se identifican, salvo la alegría, pero con una tasa de identificación muy bajo (33,33%). Esto nos puede hacer suponer que la velocidad del grupo fónico inicial no es una característica capaz de identificar las distintas emociones.
- El enfado y la sorpresa se confunden con la tristeza (con un 60,93% y 60%, respectivamente).
- La voz neutra se confunde con la alegría (54,63%).
- Al igual que en los experimentos anteriores, la tristeza nunca se confunde con la alegría, ni con la neutra, pero además en este caso:
    - La alegría nunca se confunde con el enfado.
    - El enfado nunca se confunde con la neutra.
    - La neutra nunca se confunde con el enfado.
- Las precisiones son bastante bajas en general, obteniendo valores similares para la tristeza (37,68%), que tiene una elevada tasa de identificación (86,63%), como para el enfado o la neutra (38,49% y 38,47%, respectivamente), que se confundían con la tristeza y la alegría, respectivamente. La baja precisión de la tristeza se debe a la confusión con ella del enfado y la sorpresa.

1.1.1.44.Experimentos AH, AI y AJ:

A continuación se muestran los resultados medios obtenidos en los *experimentos AH, AI* y *AJ* (considerando sólo los grupos fónicos finales):

Tabla 82. Tasas de identificación para cada emoción de los *experimentos AH, AI* y *AJ*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	22,20% (± 5,95%)	6,67%		48,90%	22,20%
Enfado	7,13%	2,37% (± 2,23%)	6,97%	70,00%	13,47%
Sorpresa	11,10%	2,23%	6,67% (± 3,57%)	75,57%	4,43%
Tristeza	2,23%	2,23%	8,90%	82,23% (± 5,47%)	4,43%
Neutro	27,77%	4,43%		46,53%	21,27% (± 5,99%)
PRECISIÓN	31,52%	13,20%	29,59%	25,44%	32,32%

Con los datos de la tabla podemos decir que:

- La única emoción que se identifica, y con una elevada tasa de identificación, es la tristeza. Pero observamos que todas las emociones se confunden con ella, por tanto, su precisión es baja (25,44%).
- La alegría y la voz neutra nunca se confunden con la sorpresa, por lo que es posible que la velocidad de los grupos finales de estas emociones presente diferencias importantes.
- La precisión de todas las emociones es bastante baja. Esto se puede deber a que la velocidad de los grupos fónicos finales no sea una característica capaz de identificar de qué emoción se trata una cierta locución, ya que se produce mucha confusión y casi todo se interpreta como sorpresa.

1.1.1.45.Experimentos AK, AL y AM:

En la siguiente tabla se muestran los resultados medios obtenidos en los *experimentos AK, AL* y *AM* (considerando sólo los grupos fónicos medios):

Tabla 83. Tasas de identificación para cada emoción de los *experimentos AK, AL* y *AM*.

	EMOCIÓN IDENTIFICADA				
EMOCIÓN INTERPRETADA	Alegría	Enfado	Sorpresa	Tristeza	Neutro
Alegría	2,57% (± 2,43%)	2,57%	2,57%	23,10%	69,20%
Enfado	5,13%		2,57%	58,33%	34,00%
Sorpresa	2,57%			56,40%	41,03%
Tristeza	2,57%	2,57%	2,57%	87,17% (± 5,14%)	5,13%
Neutro	8,33%		2,77%	16,27%	72,67% (± 7,03%)
PRECISIÓN	12,13%			36,13%	32,73%

En la tabla observamos que:

- Las mayores tasas de identificación se obtienen para la tristeza y la voz neutra.
- La alegría se confunde con la voz neutra (69,2%).
- La sorpresa nunca se identifica y se confunde principalmente con la tristeza (56,4%) y también con la neutra (41,03%).
- El enfado (al igual que la sorpresa) nunca se identifica y se confunde con la tristeza (58,33%).
- La sorpresa y la neutra nunca se confunden con el enfado.
- En este caso también son bajas las precisiones de cada emoción, debido a que se produce mucha confusión entre las distintas emociones, principalmente con la tristeza y la neutra, que son las únicas que se identifican.

7.5.3.Conclusiones de los experimentos

- La alegría obtiene una tasa de identificación mayor que la tasa de confusión con el resto de emociones sólo en el caso en el que utilizamos los grupos fónicos iniciales (obteniendo un 33,33%). Esto nos hace puede hacer pensar que la velocidad de los grupos fónicos iniciales es la que más caracteriza a la alegría.
- Cuando utilizamos los grupos fónicos medios, la confusión de la alegría con la voz neutra es elevada (69,2%), por lo que la velocidad de los grupos fónicos medios

de estas emociones será similar.

- El enfado y la sorpresa no se identifican en ningún caso, confundiéndose en todos ellos con la tristeza. La mayor tasa de confusión se da cuando usamos los grupos fónicos finales (70% y 75,57%, respectivamente).

- La tristeza es la emoción que mayor tasa de identificación tiene en todos los casos, siendo mayor cuando utilizamos los grupos fónicos medios (87,17%) y menor cuando utilizamos todos los grupos fónicos (82,2%). A pesar de ello, la mayor precisión se obtiene para este caso (en el que utilizamos todos los grupos fónicos), pero ésta es baja (47,43%), ya que el enfado y la sorpresa se confunden mucho con la tristeza.

- La voz neutra tiene la mayor tasa de identificación cuando usamos todos los grupos fónicos (74,77%), siendo muy similar a la que obtenemos cuando utilizamos los grupos fónicos medios (72,67%). Sin embargo, en los otros casos, cuando utilizamos solamente los grupos fónicos iniciales o finales, la neutra se confunde con la alegría (54,63%) y la tristeza (46,53%), respectivamente. Esto se puede deber a que la velocidad de los grupos fónicos medios sea la que caracterice la voz neutra. Para la neutra obtenemos una precisión un poco mayor que la que obteníamos para la tristeza, también cuando utilizamos todos los grupos fónicos (48,31%).

7.6.Conclusiones de los experimentos con características prosódicas

- Respecto a las características relacionadas con el contorno de F0, las conclusiones obtenidas son:

    - La emoción que mejor se reconoce es la **sorpresa**, ya que es para la que obtenemos las mayores tasas de identificación y la mayor precisión. Por tanto, podemos deducir que ésta es una emoción *prosódica*, ya que somos capaces de reconocerla empleando características relacionadas con la prosodia.

    - Las siguientes emociones que mejor se identifican son la tristeza y la voz neutra. Para estas emociones, las mayores tasas de identificación se obtienen cuando sólo utilizamos los grupos fónicos medios. Esto se puede deber a que la información prosódica relacionada con F0 que caracterice a estas emociones se encuentre en la parte central de la frase.

    - La alegría sólo se identifica cuando usamos los grupos fónicos medios y el enfado cuando usamos los grupos fónicos iniciales (pero en este caso, la alegría se confunde con el enfado). Estos resultados nos hacen pensar que la información prosódica relacionada con F0 relevante para identificar la alegría, se encuentra en la parte central de la frase; y la del enfado en la parte inicial. Puede ser que la información prosódica de la parte inicial del enfado y la alegría sea similar, y por ello, confundamos la alegría con el enfado cuando trabajamos sólo con los grupos fónicos iniciales.

    - La sorpresa nunca se confunde con la tristeza ni con la voz neutra, lo que hace que obtengamos un patrón unívoco para la sorpresa, que nos permita identificarla siempre frente a la tristeza y a la voz neutra.

    - En el único caso en el que obtenemos mejores resultados cuando no consideramos el valor medio de F0 en el vector de características es cuando usamos los grupos fónicos finales. De esto, podemos concluir que como cabía esperar, el valor medio de la F0 es un factor relevante para

identificar las emociones y que donde presenta una menor relevancia presenta es en la parte final de las frases.

- Respecto a las características relacionadas con el ritmo, las conclusiones obtenidas son:
    - La emoción para la que obtenemos una mayor tasa de identificación es la tristeza, seguida de la neutra. Pero la precisión de estas emociones es baja, debido a que el enfado y la sorpresa se confunden con la tristeza, y la alegría se confunde con la voz neutra. Por lo tanto, concluimos que la velocidad, tanto la de la locución como la de los grupos fónicos, no es una característica que nos permita reconocer una emoción frente a otra.
    - En general para todas las emociones, las mayores tasas de identificación se obtienen cuando utilizamos solamente los grupos fónicos medios.

8. CONCLUSIONES GENERALES

Una vez analizados los resultados de la aplicación de las distintas técnicas de identificación de emociones llevadas a cabo en este proyecto, es momento de extraer las principales conclusiones. Trataremos primero las conclusiones obtenidas empleando parámetros relacionados con rasgos segmentales, con datos de SES, de EMODB o combinando ambas bases de datos. A continuación, estableceremos las principales conclusiones de la identificación de emociones basada en rasgos prosódicos. También es interesante obtener conclusiones acerca de la normalización: que tipo de normalización da mejores resultados, cuando se producen mayores mejoras al emplearla, que emociones mejoran su tasa de identificación al emplear parámetros normalizados, etc. Y, finalmente, haremos una comparativa entre las conclusiones obtenidas en este proyecto y las obtenidas en estudios anteriores.

Conclusiones de la identificación de emociones basada en características segmentales sobre SES:

- Con los datos disponibles en SES hemos realizado diversos experimentos en función de los datos empleados en el entrenamiento y la clasificación. En algunos de ellos tomábamos el párrafo como unidad de clasificación y en otros las frases, pudiendo ser dependientes o independientes de texto. Los mejores resultados se obtienen para el experimento en el que la unidad de entrenamiento es la frase (tanto frases grabadas de manera independiente por el actor, como frases extraídas del párrafo cuarto de los bloques) y la unidad de clasificación es el párrafo, considerando sólo los tres primeros párrafos, consiguiendo de esta forma un experimento independiente de texto. Las posibles razones por las que puede que éste sea el experimento en el que obtengamos mejores resultados son que los modelos obtenidos son robustos, ya que están basados en dos tipos de frases con distinta naturaleza; y que la unidad de clasificación sea el párrafo, lo que hace que basemos la decisión sobre qué emoción estamos reconociendo, en un mayor

número de vectores de clasificación.

- Resulta interesante remarcar que, a priori, podríamos pensar que los resultados obtenidos en aquellos experimentos dependientes de texto fueran mejores. Pero esto no es así, lo que hace que a la hora de identificar una cierta emoción sean más relevantes las características relacionadas con rasgos segmentales, que el contenido lingüístico de la frase.

- Las emociones que mejor se reconocen son la tristeza, el enfado y la neutra; y las que peor, la alegría y la sorpresa. Éstas últimas son emociones positivas que se confunden entre ellas, y por ello baja su tasa de identificación. Aún así, en ciertos experimentos conseguimos identificarlas siempre, pero en general, son las que tienen tasas menores.

- La normalización que mejor funciona es aquella en al que estimamos la media y la varianza respecto a la voz neutra del locutor. Y la emoción para la que se obtiene la mejora más significativa al normalizar es la sorpresa.

- Las menores precisiones se obtienen para el enfado, a pesar de que sus tasas de identificación son elevadas, debido a que todas las emociones, aunque con tasas de confusión muy bajas, se confunden con él.

Conclusiones de la identificación de emociones basada en características segmentales sobre EMODB:

- Gracias a los distintos tipos de normalización aplicados, obtenemos mejoras significativas, mayores del 10%. La normalización que funciona mejor es en la que estimamos la media y la varianza respecto a la voz del locutor. Las emociones que presentan una mayor mejora al normalizar son el asco y la neutra. Adicionalmente, la precisión de todas las emociones aumenta al emplear características normalizadas, debido a que se produce una menor tasa de confusión entre las distintas emociones.

- La emoción que mejor se reconoce es la tristeza, teniendo en cuenta tanto la tasa de identificación, como la precisión; y las que peor, el miedo y la alegría.

Conclusiones de la identificación de emociones con distintos idiomas basada en características segmentales:

- Al considerar todas las emociones disponibles en ambas bases de datos, se produce una confusión de todas las emociones de EMODB, salvo la tristeza, con la sorpresa de SES.

- Es importante destacar que considerando sólo las emociones comunes a ambas bases de datos, entrenando con SES y clasificando EMODB, se obtienen tasas de identificación medias del orden de las que conseguíamos al trabajar sólo con datos de EMODB.

- La tristeza es la emoción que mejor se identifica cuando entrenamos con SES y clasificamos EMODB, y viceversa, lo que nos hace suponer que se trata de una emoción con características comunes en ambos idiomas.

- En los experimentos en los que entrenamos con SES y clasificamos EMODB se obtienen altas tasas de identificación de la alegría, pero todas las emociones se confunden con ella. Esto se puede deber a que las emociones en alemán se parezcan a la alegría interpretada por nuestro actor. Particularmente, el enfado es la emoción que más se confunde con ella, pudiendo asociar este hecho a que el enfado interpretado en alemán es un enfado en caliente, que puede tener ciertas similitudes con la alegría interpretada por el actor de SES.

- En cambio, al entrenar con EMODB y clasificar SES, las emociones tienden a confundirse con la neutra, lo que nos hace pensar que la interpretación realizada por el actor de SES sobre las distintas emociones, se parece al estado neutro de los locutores alemanes.

Conclusiones de la identificación de emociones basada en características prosódicas sobre SES:

- Considerando características relacionadas con el *contorno de F0*, conseguimos identificar con una tasa elevada la sorpresa y la tristeza. Sin embargo, la alegría y el enfado, dependiendo de los grupos fónicos que

utilicemos, se identifican o no. La alegría se identifica con los grupos fónicos medios, y el enfado con los grupos fónicos iniciales. Es decir, en principio no podemos considerar la alegría y el enfado como emociones *prosódicas*, pero si que presentan ciertas características relevantes a lo largo de una frase (en el caso de la alegría, en la parte central; y en el caso del enfado, en la parte inicial).

- El valor medio de F0 es una característica relevante a la hora de reconocer emociones, ya que al prescindir de él, se obtienen menores tasas de identificación, en general.

- Considerando características relacionadas con el *ritmo*, no conseguimos reconocer unas emociones frente a otras, ya que se produce mucha confusión. La tasa de identificación de la tristeza es elevada, pero el enfado y la sorpresa se confunden con ella.

- Los resultados obtenidos empleando rasgos segmentales son mejores que los obtenidos con rasgos prosódicos, por lo que en principio podemos pensar que las características relacionadas con rasgos segmentales nos permiten obtener unos modelos más robustos a la hora de reconocer emociones.

Conclusiones sobre la normalización de los vectores de características:

- La normalización con la que obtenemos mejores resultados, en general, es aquella en la que estimamos la media y la varianza, bien respecto a la voz del locutor o bien respecto a la voz neutra de dicho locutor.

- En los experimentos realizados con datos de EMODB se obtienen mejoras significativas al normalizar. Esto puede ser debido a que normalizando los vectores de características intentamos reducir la variabilidad presente en los datos, y en EMODB, a parte de la variabilidad del canal, está presente la variabilidad debida a la presencia de distintos locutores.

Comparativa con estudios anteriores:

Centrándonos en la base de datos castellana (SES), intentaremos establecer qué emociones podemos considerar “segmentales” y qué emociones podemos considerar “prosódicas”, para poder comparar estos resultados, con experimentos de reconocimiento automático de habla, con las conclusiones obtenidas con datos perceptuales (1.1.1.45 y1.1.1.45).

En los experimentos realizados con características relacionadas con rasgos segmentales las emociones que mejor se identifican son la tristeza y el enfado; y las que peor se identifican son la sorpresa y la alegría.

En los experimentos de identificación de emociones con características relacionadas con rasgos prosódicos relacionados con el contorno de F0 las emociones que mejor se identifican son la sorpresa y la tristeza; y las que peor se identifican son la alegría y el enfado. Respecto a los resultados obtenidos con rasgos prosódicos relacionados con el ritmo, la emoción que mejor se identifica es la tristeza, pero todas las emociones se confunden con ella.

De los dos párrafos anteriores podemos concluir que la tristeza es una emoción tanto *segmental* como *prosódica*, ya que en ambos casos conseguimos identificarla con una tasa elevada. Sin embargo, el enfado conseguimos identificarlo empleando características relacionadas con el tracto vocal, pero no con características prosódicas, lo que nos hace pensar que se trata de una emoción *segmental.* En el caso de la sorpresa, la identificamos con una tasa mayor empleando características prosódicas, pero al emplear características segmentales también conseguimos tasas altas (aunque sea junto la alegría, las emociones que peor se identifican, pero la tasa de acierto esta entre el 70% y el 90%), por tanto, podemos considerarla como una emoción tanto *segmental* como *prosódica.* La alegría es la emoción que peor identificamos en ambos casos (con características segmentales o prosódicas). Se trata de una emoción con mucha variabilidad, lo que hace que nos resulte complicado encontrar un patrón simple para reconocerla. Por ello, en este caso no podremos suponer que se trata únicamente de una emoción *segmental* o únicamente *prosódica,* sino una combinación compleja de ambas características.

GENERALES

Las conclusiones obtenidas en 1.1.1.45 fueron las siguientes: el enfado es de naturaleza *segmental* y la sorpresa de naturaleza *prosódica*. Mientras que la alegría y la tristeza son de naturaleza mixta, en parte *segmental* y en parte *prosódica*. Las conclusiones obtenidas en 1.1.1.45 son similares a las obtenidas en 1.1.1.45, salvo para el caso de la alegría. Los resultados obtenidos en 1.1.1.45 nos hacen pensar que la alegría es una emoción que no podemos considerar ni *segmental* ni *prosódica,* pero realizando experimentos en los que consideremos una combinación de ambos tipos de características, podríamos decir que se trata de una emoción tanto segmental como prosódica.

Representamos en el siguiente diagrama las conclusiones obtenidas en este proyecto comparándolas con las obtenidas en 1.1.1.45. Es importante destacar que los oyentes que evaluaron las distintas emociones en 1.1.1.45 no habían realizado ningún tipo de entrenamiento, mientras que en nuestro sistema se realiza un entrenamiento previo a la clasificación.

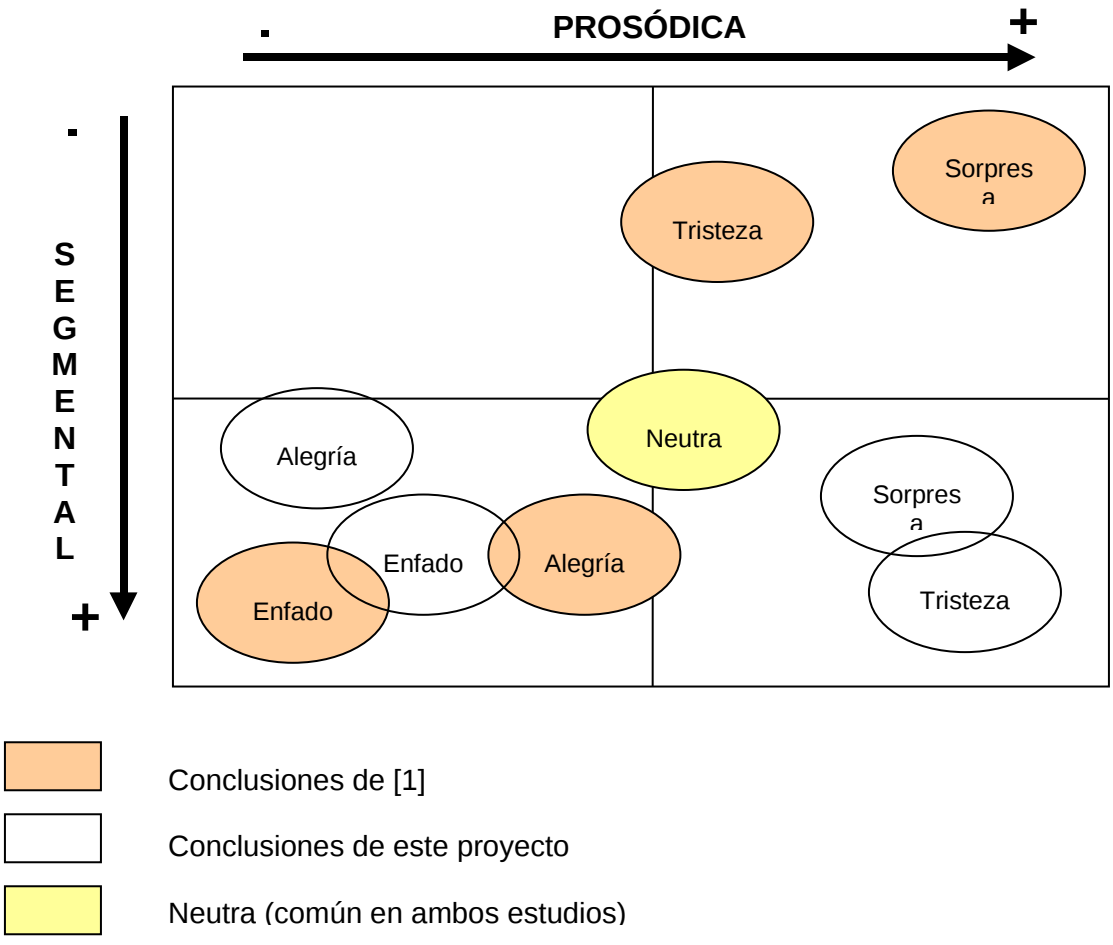


Figura 17. Comparación de las conclusiones obtenidas en este proyecto y las obtenidas en 1.1.1.45.

9. LÍNEAS FUTURAS

Tras la evaluación de los resultados de la experimentación del proyecto con las dos estrategias de identificación (basada en MFCC y prosodia), proponemos las siguientes líneas futuras de investigación:

- Aplicación de los experimentos realizados sobre la futura base de datos SEV (*Spanish Emotional Voices*). SEV dispondrá de un mayor número de datos, mayor cobertura alofónica, prosódica y emocional con relación a SES y EMODB. Además, este nuevo corpus permitirá realizar identificación de emociones multi-locutor en español.

- Realización de una serie de experimentos en los que se combinen características prosódicas y segmentales, dado que hemos observado que ciertas emociones se identifican mejor con características segmentales (como puede ser el enfado) y hay otras emociones que se identifican mejor con características prosódicas (como puede ser la sorpresa) o bien incluso identificarse con ambas fuentes de información como es el caso de la tristeza.

- Realización de una serie de experimentos en los que el vector de características tuviese las características prosódicas que permiten identificar mejor a cada una de las emociones, realizando un análisis basado en las técnicas PCA o LDA.

- Realización de experimentos de identificación de emociones basados en información prosódica con los datos de EMODB. No hemos podido realizar estos experimentos en este proyecto por no disponer de la información lingüística necesaria para dividir las frases de EMODB en sus correspondientes grupos fónicos. Este tipo de experimentos es interesante ya que se podría analizar la influencia de la existencia de distintos locutores en los rasgos prosódicos.

- Aplicación de técnicas de normalización estudiadas en el proyecto a las

características relacionadas con la prosodia; así como la evaluación de la aplicación de nuevas estrategias de normalización

- Introducción de la variación temporal en el entrenamiento para obtener los modelos de las distintas emociones, sustituyendo el modelo basado en GMM actual a modelos ocultos de Markov (HMM), estudiados en 1.1.1.45.

- Modelado de cada emoción según el conjunto de contornos de F0 de los distintos grupos fónicos del conjunto de entrenamiento de dicha emoción. La clasificación de una frase se llevaría a cabo comparando (mediante el alineamiento temporal) el contorno de F0 de los grupos fónicos de dicha frase con los contornos de F0 del modelo cada una de las emociones.

- Consideración del contorno de energía como una característica adicional relacionada con rasgos prosódicos, evaluando su relevancia sobre las bases de datos SES y EMODB.

- Otra posible característica adicional relacionada con rasgos prosódicos que podríamos evaluar es la evolución del máximo de F0 a lo largo de los distintos grupos fónicos.

- En el estudio de los modelos obtenidos para las distintas características prosódicas consideradas hemos observado que hay ciertas emociones, como la alegría, que presentan diferencias importantes en cuanto a la media y la varianza de la característica estudiada, en función del grupo fónico. Para estudiar esta variabilidad, sería interesante calcular la derivada de algunas estas características, como puede ser el máximo de F0, el rango de F0 o la velocidad de los grupos fónicos.

- Tras las conclusiones extraídas sobre la influencia de los distintos grupos fónicos en la identificación de emociones con características prosódicas, hemos observado que emociones como la alegría o el enfado, se identifican al emplear sólo los grupos fónicos medios o iniciales, respectivamente, pero no es así cuando empleamos todos los grupos fónicos. Podríamos asociar este hecho a que la utilización de una única gausiana haga que se mezclen las características de los grupos fónicos, perdiendo la información propia de cada uno de ellos y que es la

que diferencia a emociones como la alegría y el enfado, de otras. Dicho esto, la línea futura propuesta para intentar solventarlo es la utilización de un mayor número de gausianas, que permita separar las características de cada grupo fónico. En nuestro caso, por la escasez de datos, no fue posible aumentar el número de gausianas.

- Para ampliar el estudio de cómo afecta la aplicación de distintos idiomas en el reconocimiento de emociones, proponemos como línea futura la consideración de la base de datos danesa DES (Danish Emotional Speech database 1.1.1.45). De esta forma se podrán analizar las similitudes y diferencias entre el castellano, el alemán y el danés.

10.BIBLIOGRAFÍA

[1] Juan Manuel Montero Martínez. “Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano”. Tesis Doctoral, 2003.

[2] Cowie R. and Conrnelious R.R. “Describing the emotional status that are expressed in speech, in Speech Comunication”. 2003.

[3] Luengo I., Navas E., Hernáez I., Sánchez J. “Reconocimiento automático de emociones utilizando parámetros prosódicos”. 2005.

[4] Barra R., Montero J.M, Macías-Guarasa J., L.F. D’Haro, R. San-Segundo and Ricardo de Córdoba. “Prosodic and segmental rubrics in emotion identification”. IEEE Transaction on Acoustics, Speech and Signal Processing, Nº Catálogo IEEE 06CH37812:1088, mayo 2006.

[5] Javier González Domínguez. “Nuevas Técnicas de compensación de canal en reconocimiento de locutor e idioma”. 2006.

[6] Enrique G. Fernández-Abascal. María Pilar Jiménez Sánchez. María Dolores Martín Díaz. “Emoción y motivación. La adaptación humana”. Editorial Centro de Estudios Ramón Areces, S. A. Volumen I. 2003.

[7] Kleinginna, P.R., Jr. Y Kleinginna, A.M. “Motivation and Emotions - A categorized list of emotion definitions, with a suggestion for a consensual definition”. 1981.

[8] Izard C.E. “The Psychology of Emotions”. Nueva York: Plenum Press. 1991.

[9] Öhman, A., Flykt A. Y Lundqvist, D. “Cognitive neuroscience of emotion: Series affective science”. Nueva York: Oxford University Press. 2000.

[10]Fernández-Abascal, E.G. “Psicología General: Motivación y Emoción”. Madrid: centro de Estudios Ramón Areces. 1997.

[11]Chilin Shih and Greg Kochanski. “Prosody and prosodic models”. Denver Colorado. 2002.

[12]Ben Milner and Xu Shao. “Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model”. School of Information Systems, University of East Anglia, Norwich, UK. 2002.

[13]Yannis Stylianou, Olivier Cappé and Eric Moulines. “Continuous Probabilistic Transform for Voice Conversion”. IEEE Transctions of Speech and audio processing, vol. 6, n° 2, March 1998.

[14]Astrid Paeschke. “Global Trend of Fundamental Frequency in Emotional Speech”. Department of Communication Science Technical Unversity of Berlin, Germany. 2004.

[15]Berlin Database of Emotional Speech (http://pascal.kgw.tu-berlin.de/emodb/navi.html).

[16]Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. “Spoken Language Processing: A guide to Theory, Algorithm, and System Development”. Prentice Hall PTR, 1st edition. April 25, 2001.

[17]Richard O. Duda, Meter E. Hart, David G. Stork. “Pattern Classification”. Second Edition. Wiley-Interscience. 2001.

[18]R. Barra, J.M. Montero, J. Macias-Guarasa, J. Gutiérrez-Arriola, J. Ferreiros, J.M. Pardo. “On the limitations of voice conversion techniques in emotion identification tasks”. 2007.

[19]Albino Nogueiras, Asunción Moreno, Antonio Bonafonte and José B. Mariño. “Speech Emotion Recognition Using Hidden Markov Models”. Research Center TALP, Universitat Politècnica de Catalunya. Spain. Eurospeech 2001.

[20]Inger Samso Engberg, Anya Varnich Hansen. “Documentation of the Danish Emotional Speech Database”. 1996.

Carmen Rincón Llorente

Ingeniera de Telecomunicación

Abril 2007