Departamento de Teoría de la Señal y Comunicaciones

UNIVERSIDAD CARLOS III

PROYECTO FIN DE CARRERA: RECONOCIMIENTO DE EMOCIONES EN LA VOZ.

Autora: Silvia Calatrava Sierra

Tutora: Ascensión Gallardo Antolín

Leganés, abril de 2015

AGRADECIMIENTOS

Como dice el dicho popular: “Todo comienzo tiene un final” y eso es lo que representa este proyecto, el final de una etapa. Es extraño pensar que después de tanto tiempo por fin puedo cerrar esta fase de mi vida que tantas alegrías y tantas tristezas, o más bien problemas, me ha causado. Pero, por suerte para mí, no he recorrido este camino sola y he tenido a mucha gente que me ha apoyado y me ha ayudado en los peores momentos, esos en los que te planteas si no serías más feliz haciendo otra cosa, y gracias a su paciencia, consejos y a veces simplemente a su presencia he conseguido llegar donde estoy hoy. Por todo esto, y para ser justa, tengo que mencionar a un gran número de personas en estos agradecimientos.

En primer lugar, tengo que agradecer a mi madre su infinita paciencia. Debe ser verdad eso que dices de que el amor a un hijo es el sentimiento más fuerte que se puede experimentar porque me lo has demostrado una y otra vez a lo largo de los años, no sólo con tu infinita paciencia sino también con tu apoyo incondicional y por todo eso y mucho más: Muchas gracias. Por supuesto, tengo también que mencionar a mi padre en estos agradecimientos. Gracias papá por enseñarme a trabajar duro y ayudarme a convertirme en esa persona perfeccionista a la que le gustan las cosas bien hechas y gracias también por aguantar mis malos momentos en los que pensaba que nunca iba a conseguir llegar a la meta que me había marcado.

Además de mis padres, mi hermana también ha sido un pilar en el que apoyarme que ha sabido entender los “sacrificios” que he tenido que hacer y las decisiones que he tenido que tomar para llegar a este momento. Gracias por ese sentido del humor que nos ha permitido reírnos en los malos momentos y también por tu paciencia conmigo.

Gracias también a mis abuelos, tíos y primos que me han hecho sentir querida siempre y me han ayudado en todo lo que han podido y en especial a mi abuela Bernabela, a la que todos llamábamos “Bernabea”, que sin saber leer ni escribir siempre entendió que el terminar esta carrera era algo muy importante que yo tenía que hacer y que siempre hizo todo lo que estuvo en su mano para apoyarme y darme la fuerza necesaria para seguir adelante. Abuela siempre has sido un referente para mí y siempre lo serás y por todo lo vivido contigo te estoy y te estaré infinitamente agradecida.

No habría podido enfrentarme a este proyecto sin la ayuda de mi tutora, Ascen. Muchas gracias por estar siempre dispuesta a ayudarme con cualquier duda y por la paciencia que has tenido con el proceso de escritura del mismo que al final se ha hecho realmente largo.

También quiero mencionar a mis amigos de dentro y fuera de la universidad que han estado siempre ahí cuando lo he necesitado. Zaida, tú has sido esa amiga con la que he crecido y con la que he aprendido a enfrentarme al mundo. Laura, Raquel y Jorge que aunque entre nosotros nos llamemos los Nadies sois unos Álguienes muy importantes para mí. Qué habría sido de mí sin nuestras tardes de biblioteca y sin nuestros momentos de risas y quejas contra el mundo, por todo esto y mucho más que no hace falta comentar muchas gracias a los cuatro.

Paula y Marisa que aunque lejos físicamente siempre han estado a mi lado dándome su apoyo. Ruth, Lucía y Nacho, creo que sois lo mejor que he sacado de todos estos años en la carrera. Me habéis demostrado que aunque hay mucha gente oportunista hay otra mucha que estará cerca cuando se la necesite.

Natalia e Irene, ¿quién nos iba a decir cuando nos conocimos en el PIC de Leganés que acabaríamos siendo tan buenas amigas? Gracias por todos los buenos momentos que hemos pasado juntas y por los que seguiremos compartiendo.

Y ya por último, gracias a mis amigos de la infancia y del instituto (Miguel, Patri, Lidia, Sara) y a toda la gente que conocí como becaria en el PIC de Leganés (Mari Carmen, Óscar, Mamen, Antonio, Raúl, Sheila, etc.) que tanto me enseñaron y al resto de compañeros de esos años a los que espero seguir viendo al menos una vez al año.

Muchas gracias.

Silvia.

RESUMEN

El reconocimiento de emociones en el habla es un problema que puede abordarse desde distintos frentes. Por una parte, es necesario elegir un sistema de reconocimiento de emociones que se adapte a nuestras necesidades. Por otro lado, la elección de las características acústicas de las muestras de voz incluidas en el proceso, así como los métodos utilizados para la extracción de las mismas es otro de los puntos críticos del reconocimiento de emociones.

La finalidad de este proyecto es la de obtener una serie de conclusiones con respecto a la utilización de distintos conjuntos de características acústicas en el proceso de reconocimiento de emociones.

En este trabajo, una vez seleccionadas las características, así como las bases de datos a partir de cuyas muestras se obtienen dichas características, se han realizado un conjunto de experimentos a partir de cuya observación y comparación se han extraído una serie de conclusiones que podrían generar futuras líneas de investigación.

ABSTRACT

Emotion recognition in speech is a nowadays problem which can be approached from two different perspectives. On one hand, it is necessary to choose an emotion recognition system which can give us the best possible results. On the other hand, the acoustic features which are going to be used are another important factor, just as important as the methods used to obtain these characteristics which are going to be used in the recognition process.

The aim of this Project is to obtain a set of conclusions with respect to the use of a subset of different acoustic features in the emotion recognition process.

In this work, once this subset of features and the two databases which are going to be used to extract them have been selected, a large number of experiments have been performed. From the achieved results, a set of conclusions which can lead to further research have been drawn.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

Para comenzar este proyecto fin de carrera, me gustaría resaltar la creciente importancia que está adquiriendo la investigación acerca del reconocimiento de emociones en la voz humana para el futuro desarrollo de sistemas que pudieran orientarse, no sólo hacia el público más habitual, sino también hacia un público con diversidad funcional como puede ser el colectivo invidente.

Existen diversas motivaciones para llevar a cabo la identificación de emociones en el habla. Por ejemplo, en una interacción persona-máquina, como puede ser aquella que se realiza con los teléfonos de atención al cliente de los principales operadores, nuestra máquina puede generar respuestas más apropiadas conociendo el estado emocional de su interlocutor. Es decir, dependiendo de la respuesta que dicha máquina haya obtenido de su interlocutor, así como de su estado emocional, ésta debe ser capaz de cambiar de estrategia para obtener unos resultados más satisfactorios tanto para el usuario, produciendo una sensación de satisfacción al ser comprendido más rápidamente, como para la máquina, al obtener la información con menor dificultad y mayor rapidez.

Este proyecto se propuso con el fin de obtener una serie de conclusiones con respecto al conjunto de características acústicas utilizadas en los procesos de reconocimiento de emociones y realizar, de esta forma, una serie de aportaciones a este problema. Por tanto, el objetivo de este proyecto se basa en extraer, a partir de los diversos experimentos detallados en él, un conjunto de conclusiones sobre qué características o parámetros acústicos producen mejores resultados en dichos sistemas.

Para llevar a cabo este proyecto, primero fue necesario realizar un estado del arte sobre dicho tema para conocer más en profundidad qué parámetros y técnicas, así como bases de datos y emociones, se utilizan en el proceso de reconocimiento de emociones y reunir de esta forma información que pudiese resultarnos útil. Este estado del arte se recoge en el segundo capítulo de este documento: ESTADO DEL ARTE.

Una vez finalizado dicho estudio, se estableció qué características iban a ser incluidas en el conjunto de experimentos que se llevarían a cabo en este caso. El número de parámetros acústicos diferentes que pueden incluirse, o no, en un experimento es muy elevado, por lo que para obtener unas conclusiones coherentes escogimos un subconjunto de estos. Las diferentes características incluidas en los diferentes experimentos, así como los procesos de extracción de las mismas se recogen en el tercer capítulo de este proyecto: EXTRACCIÓN DE LAS CARACTERÍSTICAS ACÚSTICAS PARA RECONOCIMIENTO DE EMOCIONES.

Otra decisión muy importante a la hora de llevar a cabo esta serie de experimentos es la de establecer qué sistema de reconocimiento iba a ser utilizado, puesto que el sistema así como el procedimiento a seguir en cada experimento debía mantenerse constante en todos ellos para que el cambio producido en los resultados obtenidos fuese debido únicamente a la diferencia

de parámetros acústicos utilizados en un experimento determinado. El sistema de reconocimiento utilizado además de las bases de datos empleadas en estos experimentos se encuentran explicados en el cuarto apartado de este documento: SISTEMA DE RECONOCIMIENTO DE EMOCIONES Y BASE DE DATOS.

En el capítulo número cinco, RESULTADOS EXPERIMENTALES, se lleva a cabo una descripción detallada de los resultados obtenidos a partir de las muestras de las dos bases de datos utilizadas, además de la descripción de los experimentos llevados a cabo para la obtención de dichos resultados.

Por último, las conclusiones extraídas de dichos experimentos se resumirán en el sexto capítulo, en el que se incluirán además unas posibles líneas futuras de investigación: CONCLUSIONES Y LÍNEAS FUTURAS.

2. ESTADO DEL ARTE

Introducción

Para comenzar este estado del arte acerca del proceso de reconocimiento de emociones en la voz, es necesario adquirir o refrescar una serie de conocimientos básicos sobre los aspectos psicológicos, biológicos y lingüísticos de las emociones.

Desde el punto de vista psicológico, diversas ideas han ido surgiendo acerca de las emociones a lo largo de la historia del ser humano. En la Edad Moderna, Descartes introdujo la idea de que bajo todo el conjunto de la vida emocional subyace una agrupación primaria de emociones, mientras que Darwin, ya en la Edad Contemporánea, estableció la idea de que las emociones son inseparables de las costumbres más prácticas o duraderas, aquellas seleccionadas por la evolución debido a su valor para la supervivencia. A su vez, William James describió las emociones como la percepción de la mente de condiciones fisiológicas que aparecen como respuesta a un estímulo y Magda B. Arnold, por su parte, enfatizó que las emociones contienen una valoración cognitiva que alerta al organismo de situaciones que pueden tener algún tipo de significado especial. [1]

Sin embargo, es posible que los aspectos más interesantes provengan de la característica de causa y efecto de las mismas. El espacio de activación y evaluación de las emociones considera el estímulo que excita la emoción, la habilidad cognitiva para evaluar la naturaleza del estímulo y su respuesta física y mental al mismo. La respuesta mental al estímulo se traduce en un estado emocional concreto, mientras que la respuesta física se traduce en un estado de “lucha o huida”, “fight or flight”. Cuando percibimos una amenaza física o psicológica, una parte de nuestro cerebro, llamada hipotálamo, se activa, segregando una hormona llamada CFR. Esto estimula la secreción de una serie de hormonas e impulsos eléctricos que causarán la liberación de Norepinefrina y Adrenalina provocando un incremento en la frecuencia cardiaca, una aceleración de la respiración y un aumento de la presión arterial y el nivel en sangre, así como en la actividad muscular entre otros cambios fisiológicos. Todo esto prepara nuestro cuerpo para hacer frente a la amenaza enfrentándonos, luchando, o huyendo de ella.

Desde un punto de vista biológico, las respuestas físicas y emocionales pueden considerarse como patrones seleccionados evolutivamente debido a su valor para la supervivencia. Como explicamos anteriormente, las emociones tienen un efecto sobre diversos aspectos de nuestro cuerpo, como es la temperatura, el pulso o la actividad muscular. Las amígdalas, por ejemplo, reciben diversos estímulos y participan en tareas de aprendizaje sobre las recompensas que estos nos proporcionan, mientras que la corteza orbitofrontal está involucrada en la preparación de respuestas ante estos estímulos. El resultado de todos estos cambios hace que el estado emocional de una persona se manifieste en sus palabras y sus expresiones faciales.

En el aspecto lingüístico, la elección de las etiquetas apropiadas para el reconocimiento de las emociones es muy importante. Puesto que el número de emociones del ser humano es gigantesco, es necesario hacer una división entre emociones primarias y secundarias. Las emociones primarias que se utilizan habitualmente para realizar el reconocimiento son: alegría, tristeza, miedo, enfado, sorpresa y disgusto o asco. [2] Estas emociones primarias también reciben el nombre de emociones básicas y son más primitivas y universales que todas las demás.

El reconocimiento de emociones en la voz humana no es algo novedoso. Los primeros estudios se realizaron a mediados de los años ochenta utilizando las propiedades estadísticas de algunas características acústicas. Diez años después, la evolución de los ordenadores hizo posible la implementación de algoritmos de reconocimiento de emociones más complicados donde las características de la voz podían ser estimadas de forma más precisa. En la actualidad, los estudios se centran en encontrar nuevas combinaciones de clasificadores que aumenten la eficiencia de estas clasificaciones en aplicaciones de tiempo real.

Para finalizar esta introducción sólo nos cabe mencionar que las emociones pueden identificarse a través de las expresiones faciales y del habla, además de otros aspectos biológicos, algunos de los cuales han sido nombrados anteriormente.

En este estado del arte citaremos algunas de las características de la voz que pueden resultar más importantes en el reconocimiento de emociones, así como algunos de los procesos utilizados para obtenerlas. También mencionaremos algunas de las bases de datos más utilizadas en los estudios actuales y de los métodos más empleados para la clasificación de las emociones primarias. Por último, nos centraremos, en el reconocimiento de emociones en el habla, intentando llevarlo a cabo independientemente del idioma en el que se encuentran grabadas las muestras de las bases de datos y de las frases utilizadas para mostrar dichas emociones, realizando distinciones entre los estudios realizados con muestras obtenidas a partir de una sola base y los realizados con una mezcla de muestras de diversas bases.

Características del discurso emocional.

Para llevar a cabo el reconocimiento de emociones en el habla, es muy importante realizar una correcta identificación de los rasgos paralingüísticos que representan el estado emocional del hablante. Las variables de la voz que pueden ayudarnos a identificar los diferentes estados emocionales del hablante son, entre otras, la frecuencia fundamental junto con su contorno, la distribución de la energía en el espectro de la voz y la velocidad de la voz o tasa de habla.

En esta sección trataremos de abordar las principales características a través de las cuales podemos llevar a cabo este reconocimiento de emociones, explicando sus variaciones en función del tipo de emoción que se representa, así como los principales métodos utilizados para estimar estas características.

Al trabajar con voz humana, se puede observar que las propiedades de las señales de voz cambian lentamente. Debido a esto, la obtención de las características que van a ser utilizadas más adelante para el reconocimiento de emociones se realiza utilizando una ventana de la siguiente forma:

$$f_s(n;m) = s(n)w(m-n) \quad [3]$$

Donde $s(n)$ es la señal de voz y $w(n-m)$ es la ventana de longitud N_w. La longitud de la ventana tiene que ser suficientemente pequeña para que las propiedades de interés no cambien y suficientemente grande para poder estimar los parámetros.

A continuación, vamos a proceder a explicar algunos de los métodos más utilizados para obtener estas características.

Pitch o frecuencia fundamental

La señal de la frecuencia fundamental engloba no sólo esta frecuencia, sino también el contorno de la misma que se utiliza para describir las variaciones del pitch en términos de patrones geométricos.

La señal del pitch contiene información acerca de las emociones, puesto que depende de la tensión de las cuerdas vocales además de la presión del aire procedente de la cavidad pulmonar y es producida por la vibración de las cuerdas vocales. A partir de esta señal obtenemos las dos características que se utilizan principalmente en los estudios sobre la voz humana: la frecuencia fundamental (*pitch*) y la velocidad de vibración del aire a través de la glotis (*glottal air velocity o glottal volumen velocity*).

El tono o frecuencia fundamental de la voz se define como la tasa de vibración de las cuerdas vocales, mientras que a la velocidad de vibración de la glotis denota la velocidad a la que el aire atraviesa la glotis durante la vibración de las cuerdas vocales.

El primero de los métodos que van a ser descritos en este estado del arte se basa en el cálculo de la autocorrelación (*autocorrelation of center-clipped frames* [3]).

Una vez obtenida la señal de la base de datos utilizada, esta es filtrada con un filtro paso bajo centrado en 900 Hz y dividida en ventanas tal y como se ha explicado anteriormente; $f_s(n;m)$. A continuación, se procede a aplicar el recorte o clipping a cada una de estas ventanas. Esta técnica es un procedimiento no lineal que previene que el primer formante que encontramos en la ventana interfiera con la frecuencia fundamental. Para aplicar este procedimiento se fija un umbral, $C_{t/r}$, al 30% del valor máximo de la ventana.

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{t/r}, & |f_s(n;m)| > C_{t/r} \\ 0, & |f_s(n;m)| < C_{t/r} \end{cases}$$

El siguiente paso de este método, consiste en el cálculo de la autocorrelación de la ventana recortada en la que η representa el intervalo o el retardo.

$$r_s(\eta;m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n;m)\hat{f}_s(n-\eta;m)$$

Por último, se obtiene la frecuencia fundamental de la ventana estableciendo F_s como frecuencia de muestreo y F_l y $F_{/l}$ como las frecuencias de pitch más baja y más alta, respectivamente, que pueden ser percibidas por el ser humano.

$$\widehat{F_0}(m) = \frac{F_s}{N_w} \underset{\eta}{\operatorname{argmax}}\{|r(\eta;m)|\}_{\eta=N_w(F_l/F_s)}^{\eta=N_w(F_{/l}/F_s)}$$

Para determinar la velocidad de vibración de la glotis se utiliza el valor máximo de la autocorrelación.

$$\max_{\eta}\{|r(\eta;m)|\}_{\eta=N_w(F_l/F_s)}^{\eta=N_w(F_{/l}/F_s)}$$

El segundo método utilizado para estimar el pitch se basa en la transformada wavelet. La extracción del periodo del pitch se basa en el doble pase de la transformada wavelet dual. En la transformada wavelet dual b representa el índice de tiempo, 2^j es un factor de escalado, $s(n)$ la señal de voz muestreada y $\emptyset(n)$ un spline cúbico wavelet. La transformada wavelet dual representa la convolución de la señal de voz con la señal wavelet abatida en el dominio del tiempo (*time-reversed wavelet*). Este proceso se repite para tres escalas diferentes de wavelets.

$$D_yWT\left(b,2^j\right) = \frac{1}{2^j} \sum_{n=-\infty}^{+\infty} s(n)\emptyset\left(\frac{n-b}{2^j}\right)$$

En el primer pase, el resultado de la transformada se enventana con una ventana rectangular de 16 ms y un solapamiento de 8 ms. La frecuencia fundamental se calcula estimando el máximo de $D_yWT\left(b,2^j\right)$ de las tres escalas.

El segundo paso se incluyó para marcar el periodo de excitación del pitch (*pitch epoch*) en el discurso exaltado (*stressed speech*). En este segundo paso, la misma transformada wavelet se aplica únicamente a los intervalos de los periodos de pitch del primer paso que tengan un pitch

epoch mayor que el 150% del valor medio de los pitch epochs medidos durante el primer paso. El resultado de esta segunda transformada wavelet se enventana con una ventana de 8 ms con 4 ms de solapamiento para capturar los pitch epochs que se producen en el discurso exaltado.

Otro método utilizado para estimar el pitch es, por ejemplo, el algoritmo PRAAT, el cual está también basado en la autocorrelación.

Operador de energía de Teager

Otra característica importante es el número de armónicos producidos por el flujo de aire al pasar por el tracto vocal. A continuación, un método para calcular este número de armónicos en la señal de voz es descrito.

Suponiendo que una ventana de la señal de voz, $f_s(n;m)$, tiene un simple armónico que puede ser considerado como una señal sinusoidal AM-FM, ésta puede ser representada en tiempo discreto de la siguiente forma:

$$f_s(n;m) = \alpha(n;m)\cos\big(\Phi(n;m)\big) = \alpha(n;m)cos\left(\omega_c n + \omega_{\not h}\int_0^n q(k)dk + \theta\right)$$

En esta representación podemos diferenciar $\alpha(n;m)$ como amplitud instantánea y $\omega_i(n;m)$ como frecuencia instantánea. Además, ω_c es la frecuencia de portadora, $\omega_{\not h} \in [0, \omega_c]$ es la máxima desviación de la frecuencia y θ es una constante de fase.

$$\omega_i(n;m) = \frac{d\Phi(n;m)}{dn} = \omega_c + \omega_{\not h} q(n), \quad |q(n)| \leq 1$$

Para estimar estos dos parámetros se puede utilizar el operador de energía de Teager (TEO):

$$\Psi[f_s(n;m)] = (f_s(n;m))^2 - f_s(n+1;m)f_s(n-1;m)$$

Al aplicar este operador sobre la señal AM-FM sinusoidal se obtiene el cuadrado del producto de sus componentes.

$$\Psi[f_s(n;m)] = \alpha^2(n;m)\sin\left(\omega_i^2(n;m)\right)$$

Una vez calculado el operador, la amplitud instantánea así como la frecuencia instantánea pueden ser estimadas, suponiendo $\Delta_2 = f_s(n+1;m) - f_s(n-1;m)$, de la siguiente manera:

$$\omega_i(n;m) \approx \arcsin\left(\sqrt{\frac{\Psi[\Delta_2]}{4\Psi[f_s(n;m)]}}\right)$$

$$\alpha(n;m) \approx \frac{2\Psi[f_s(n;m)]}{\sqrt{\Psi[\Delta_2]}}$$

Este procedimiento se basa en la suposición de que en cada ventana de la señal de voz, cada uno de los armónicos tiene una sola amplitud y frecuencia instantáneas. Si esta ventana tiene un único armónico, el resultado del operador de Teager, $\Psi[f_s(n;m)]$, será una constante, mientras que en el caso de tener más de un armónico este operador dependerá de n.

Otra característica que puede ser utilizada en el reconocimiento de emociones es el conjunto de coeficientes del polinomio que describe la envolvente de la autocorrelación del TEO.

Formantes

Los formantes son una de las características que representan el tracto vocal, para ser más específicos, representan las resonancias del mismo.

La localización de estos formantes en el dominio de la frecuencia, depende de la forma y las dimensiones del tracto vocal. Cada uno de estos formantes se caracteriza por su frecuencia central y su ancho de banda. A continuación un método para estimar estos dos parámetros será descrito.

Un método sencillo, utilizado para estimar los formantes que aparecen en cada una de las ventanas de la señal de voz, es aquel basado en el análisis de predicción lineal.

En este método, consideramos que el tracto vocal está modelado por un polinomio de orden M que únicamente contiene polos y en el que los coeficientes $\hat{a}(i)$ sean coeficientes de predicción lineales (*linear prediction coefficients* o *LPCs*).

$$\hat{\theta}(z) = \frac{1}{1 - \sum_{i=1}^{M} \hat{a}(i) z^{-i}}$$

Los ángulos de los polos de $\hat{\theta}(z)$ que se encuentran más lejos del origen en el plano Z son indicadores de las frecuencias de los formantes. Estudios realizados han demostrado que los dos primeros formantes son los más afectados por los estados emocionales del hablante.

Este método plantea el problema de la falsa identificación de formantes, por lo que, para solucionarlo, se propuso una nueva aproximación para realizar esta estimación.

En esta nueva aproximación se utiliza la frecuencia instantánea obtenida a partir del operador de energía de Teager (*TEO*) como estimación aproximada de la situación de un formante para, iterativamente, refinar la frecuencia central del formante. Para ello, utilizaremos la frecuencia central obtenida a través del método de LPCs, en este caso esta frecuencia se representará con $f_c^l(m)$.

Para aislar la banda que contiene el formante se hará pasar la señal a través de un filtro paso banda cuyo ancho de banda será representado por β.

$$G_1(n) = \exp[-(\beta n T)^2]\cos(2\pi f_c^l T n)$$

En esta ecuación, $G_1(n)$ representa la respuesta impulsional de un filtro de Gabor paso banda (un filtro de Gabor es un filtro lineal cuya respuesta impulsional es una función sinusoidal multiplicada por una función gaussiana), mientras que T representa el periodo de muestreo.

El siguiente paso consiste en estimar $f_c^{l+1}(m)$ a partir de la siguiente aproximación:

$$f_c^{l+1}(m) = \frac{1}{2\pi N_w} \sum_{n=m-N_w+1}^{m} \omega_i(n;m)$$

En esta ecuación, $f_c^{l+1}(m)$ es la frecuencia central del formante durante la iteración $l+1$.

Al tratarse de un algoritmo iterativo, cuando la distancia entre $f_c^{l}(m)$ y $f_c^{l+1}(m)$ sea menor que 10 Hz el método terminará y $f_c^{l+1}(m)$ será la frecuencia estimada del formante.

Si el criterio de parada anterior no se satisface, la señal se filtra otra vez con el filtro de Gabor, centrado ahora en $f_c^{l+1}(m)$, se calcula de nuevo la frecuencia central con la aproximación y se vuelve a comprobar el criterio de parada. Este método finaliza tras unas pocas iteraciones.

Secciones transversales del tracto vocal

Las secciones transversales del tracto vocal se ven representadas por un modelo de multitubos (*multi-tubes*) sin pérdidas. Cada uno de los tubos está definido por su sección transversal y su longitud, además asumimos que en ellos no hay ningún tipo de pérdida de energía. Aunque este tipo de modelo es más realista cuanto mayor número de tubos son considerados, por simplicidad en los cálculos, se tiende a suponer 10 secciones transversales de longitud determinada.

El área de la sección transversal más cercana a la glotis es representada por A_1 y las siguientes, hasta los labios, son representadas por A_2, A_3, … y así secuencialmente.

Energía de la señal de voz

La energía de las diferentes ventanas de la señal de voz puede ser utilizada para el reconocimiento de emociones, puesto que está relacionada con el nivel de excitación de éstas.

$$E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_s(n;m)|^2$$

Log frequency speech power coefficients (LFPC)

Los coeficientes LFPC describen la forma del espectro de la señal de voz. Sus componentes ofrecen una medida de la distribución de potencia en las diferentes sub-bandas del dominio de la frecuencia.

Para calcular estos coeficientes, cada uno de los segmentos de la señal previamente obtenidos, son enventanados otra vez, pero esta vez empleando para ello una ventana de Hamming que nos permite aplicar a cada muestra un peso distinto además de reducir el ensanchamiento del espectro (*spectral leakage*).

Una vez realizado el proceso anterior, esta señal se transforma al dominio de la frecuencia a partir de la transformada discreta de Fourier (*DFT*) y sus componentes se separan utilizando para ello 12 filtros paso banda, cuyas frecuencias centrales y anchos de banda han sido calculados a partir de las siguientes ecuaciones.

$$b_1 = C$$

$$b_i = \alpha b_{i-1}, \quad 2 \leq i \leq 12$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2}$$

En estas ecuaciones C es el ancho de banda del primer filtro y f_1 su frecuencia central, mientras que α es el factor de crecimiento logarítmico.

La salida n-ésima del banco de filtros viene dada por:

$$S_t(n) = \sum_{k=f_n-(b_n/2)}^{f_n+(b_n/2)} (X_t(k)W_n(k))^2, \qquad n = 1, 2, \dots, 12$$

En la ecuación anterior reconocemos $X_t(k)$ como la componente k-ésima del espectro de la señal enventanada y $W_n(k)$ como la respuesta en frecuencia de uno de los filtros anteriores en la que l_n y h_n son los límites inferior y superior del n-ésimo filtro.

$$W_n(k) = \begin{cases} 1 & l_n \leq k \leq h_n \\ 0 & en\ otro\ caso \end{cases} \qquad n = 1, 2, \dots, 12$$

Los LFPCs se ven representados por los parámetros $SE_t(n)$ en los que N_n es el número de componentes espectrales pertenecientes al n-ésimo filtro. Estos parámetros indican la distribución de energía entre las diferentes bandas.

$$SE_t(n) = \frac{10 \cdot log_{10}(S_t(n))}{N_n}$$

Por último, simplemente mencionar que los parámetros MFCC, Fflfbe y LPCC son también muy utilizados en el proceso de reconocimiento de emociones. Los dos primeros serán explicados con más detalle en las siguientes secciones por tratarse de los parámetros utilizados para realizar este proyecto fin de carrera.

Conjunto de emociones.

Pese a que en los diferentes estudios llevados a cabo no se utiliza exactamente el mismo conjunto de emociones, podemos reconocer al menos seis emociones que intervienen en la mayor parte de éstos.

La ira, la sorpresa, la alegría, el miedo, el asco y la tristeza son un pequeño conjunto de emociones que se utiliza en este tipo de estudios. Todas ellas pertenecen al conjunto primario, o básico, de emociones del que hablábamos en la introducción de este estado del arte. En algunos estudios la neutralidad, así como el aburrimiento, son también empleados para este reconocimiento.

En este apartado trataremos de explicar a grandes rasgos como varían algunos de los diferentes parámetros que representan las características de la voz en función de la emoción que predomine. Los resultados que se muestran a continuación proceden del artículo *Speech emotion recognition using hidden Markov models* [2].

Emociones	Ira	Sorpresa	Alegría
Contorno del pitch	Sigue una forma regular a excepción de las sílabas tónicas. Estas sílabas producen irregularidades (saltos), ascienden rítmica y frecuentemente.	Se desliza de repente hacia un nivel superior en las sílabas tónicas para después bajar a un nivel intermedio o bajo en la última sílaba.	Tiene una línea descendente, asciende frecuentemente a intervalos irregulares.
Frecuencia central media	Aumenta su media.	-	Aumenta su media.
Rango del pitch	Muy amplio.	Amplio.	Muy amplio.
Intensidad	Alta.	-	Alta.
Velocidad (Rate)	Alta.	Normal.	Alta.
Espectro	Punto medio alto para espectros medios y partes no fricativas.	-	Aumenta en altas frecuencias.
Calidad de la voz (*Quality of voice*)	Tensa y entrecortada. Tono ronco.	Entrecortada.	Tensa y entrecortada. Tono estruendoso.

Tabla 1. Características acústicas de las emociones: Ira, Sorpresa y Alegría.

Emociones	Miedo	Asco	Tristeza
Contorno del pitch	Desintegración del modelo y gran número de cambios en la dirección del pitch	Ancho, variación hacia abajo al final.	Variaciones hacia abajo.
Frecuencia central media	Aumenta su media	Mucho más bajo.	Por debajo de la media.
Rango del pitch	Aumenta su rango.	Un poco más amplio.	Un poco más estrecho.
Intensidad	Normal.	Baja.	Disminuye.
Velocidad (Rate)	Alta.	Mucho más rápida.	Un poco más lenta.
Espectro	Aumenta en altas frecuencias.	-	Tiene variaciones hacia abajo.
Calidad de la voz (Quality of voice)	Tensa, sonoridad irregular.	Tono sonoro, que retumba.	Relajadas, resonante.

Tabla 2. Características acústicas de las emociones: Miedo, Asco y Tristeza.

Otros estudios demuestran que la ira es la emoción con mayor energía y frecuencia central, mientras que el asco se expresa con un pitch y una intensidad bajos y una velocidad de discurso menor que la de la emoción neutra [3]. El miedo, sin embargo, conlleva un nivel de frecuencia central alto, así como un incremento en la intensidad, mientras que para la tristeza tanto la intensidad media como la frecuencia central tienen niveles bajos.

El contorno del pitch, nos permite separar el miedo de la alegría. Este parámetro es similar en el miedo y la tristeza puesto que en ambos casos lo que obtenemos, aproximadamente, es una pendiente hacia abajo a diferencia de la alegría en la que esta pendiente es hacia arriba.

Todas estas conclusiones se ven reflejadas en la tabla que aparece a continuación.

	Pitch				Intensidad		Tiempo	
	Media	Rango	Varianza	Contorno	Media	Rango	Velocidad	Duración
Ira	>>	>	>>		>>H, >M	>	<H, >M	<
Asco	<	>H, <M			<		<<H, <M	
Miedo	>>	>		↗	=>			<
Alegría	>	>	>	↘	>	>		<
Tristeza	<	<	<	↗	<	<	>H, <M	>

Tabla 3. Conclusiones sobre las características acústicas de las emociones.

El significado de los símbolos de la tabla anterior es el siguiente: <: disminuye, >: aumenta, =: no varía respecto del neutral, ↗ : sube, ↘ : baja, H: hombre y M: mujer.

Bases de datos.

Las diferentes muestras de voz usadas en los diversos estudios proceden, en su mayor parte, de una serie de base datos recopiladas para ayudar a los investigadores en su trabajo. Hoy en día hay un gran número de bases de datos que ofrecen muestras grabadas en estudios o en condiciones más realistas, muestras espontáneas del hablante o de actores profesionales e incluso bases en las que se ha grabado la interacción de personas adultas con bebés o niños muy pequeños.

En esta sección, citaremos algunas de las bases de datos más utilizadas así como algunas de sus principales características.

Danish Emotional Speech (*DES*)

Los datos obtenidos a partir de esta base de datos contienen diversas frases neutras en danés grabadas por actores profesionales, en este caso 2 hombres y 2 mujeres, que expresan cinco estados emocionales: ira, alegría, neutral, tristeza y sorpresa. Las grabaciones se hicieron con una precisión de 16 bits y una frecuencia de muestreo de 20 kHz.

Berlin Emotional Speech Database (*EMO-DB*)

Las muestras en alemán contenidas en EMO-DB han sido grabadas en un estudio por actores y actrices profesionales, 5 hombres y 5 mujeres. Las emociones expresadas en este conjunto de frases neutras son: ira, aburrimiento, asco, miedo, alegría, neutralidad y tristeza. Estas muestras se tomaron con una precisión de 16 bits al igual que en la base DES y una frecuencia de muestreo de 22 kHz.

eNTERFACE

Esta base de datos contiene 1277 muestras de emociones audiovisuales públicas consistentes en ira, asco, miedo, alegría, tristeza y sorpresa inducidas. eNTERFACE está compuesta por un conjunto de frases predefinidas en inglés, grabadas en un ambiente de oficina. Cuarenta y dos individuos de catorce países distintos intervinieron en esta grabación. Cada una de estas personas escuchaba seis historias cortas consecutivas y cada una de estas historias inducía una emoción particular.

Una vez terminado el proceso de escucha, los individuos tenían que reaccionar a cada situación pronunciando frases que previamente habían leído y encajaban con las historias. A cada emoción le corresponden cinco frases.

Airplane Behaviour Corpus

Se trata también de una base de datos audiovisual creada para utilizarse para aplicaciones de vigilancia en transportes públicos. Para inducir una determinada emoción se utilizó un script que llevaba a los sujetos a través de una historia guiada. El marco general de la historia englobaba unas vacaciones con vuelo de ida y vuelta en la que se incluían escenas en las que se servía comida errónea, en las que había turbulencias, se quedaban dormidos o hablaban con un vecino, entre otros.

Esta base de datos fue grabada por 8 sujetos de entre 25 y 48 años en alemán.

Speech Under Simulated and Actual Stress (*SUSAS*)

La base SUSAS es la primera referencia de grabaciones espontáneas. Las muestras que aparecen aquí están grabadas en condiciones de miedo y estrés en la que las muestras se encuentran parcialmente ocultas por el ruido. Estos 7 individuos, 3 mujeres y 4 hombres, fueron grabados en una montaña rusa o en caída libre, entre otras situaciones. Los estados emocionales recogidos en las expresiones predefinidas en inglés de esta base son: neutralidad, miedo, estrés medio, estrés alto y gritando.

Audiovisual Interest Corpus (*AVIC*)

AVIC es una base de muestras espontáneas de emociones en la que el contenido hablado no está restringido. Para componer esta base, un vendedor de productos muestra a un sujeto de entre 21 una presentación comercial en inglés. El nivel de interés es anotado como interacción aburrida, neutral o alegre.

Sensitive Artificial Listener (*SAL*)

Esta base de Belfast forma parte de la base de datos HUMAINE. En ella se encuentran grabaciones audiovisuales de conversaciones naturales entre un humano y un ordenador, utilizando para ello una interfaz diseñada para permitir que los usuarios trabajasen con un rango de estados emocionales.

SmartKom

SmartKom contiene un conjunto de muestras de discurso espontáneo y emociones naturales consistente en diálogos de Wizard-Of-Oz (el Mago de Oz) en alemán e inglés en los que el ruido se superpone a las grabaciones. Los estados emocionales que se representan en esta base de datos son: neutralidad, alegría, ira, impotencia, reflexión y sorpresa entre otros.

Vera-Am-Mittag (*VAM*)

La base VAM está compuesta por grabaciones audiovisuales de programas de entrevistas alemanes en los que las discusiones son auténticas y para las que no se ha utilizado ningún tipo de script. Los temas tratados son en su mayoría temas personales tales como relaciones amorosas o preguntas sobre la paternidad o maternidad del invitado y el lenguaje utilizado es coloquial cubriendo diferentes dialectos alemanes.

Kismet

Esta base de datos contiene un total de 1002 muestras de inglés americano en las que tres mujeres se comunican con niños pequeños o bebés. Estas muestras, grabadas en diversas condiciones de ruido, están divididas en cinco clases: aprobación, atención, prohibición débil, efecto calmante (soothing) y neutral. Estas grabaciones tienen 16 bits por muestra y su frecuencia de muestreo varía entre 8 y 22 kHz.

BabyEars

BabyEars contiene 509 grabaciones en inglés americano de 6 padres y 6 madres interactuando con sus hijos. Las expresiones utilizadas son naturales y nada exageradas. Tres clases de emociones se tienen en cuenta en esta base: aprobación, atención y prohibición.

Técnicas de Clasificación

Una vez se han obtenido las diferentes características de la ventana de voz, se procede a la clasificación de las mismas. El resultado de esta clasificación es un valor, una etiqueta, que representa el estado emocional de una de las muestras.

Para evaluar estas técnicas de clasificación se utiliza el método de validación cruzada (*cross-validation*), en el que las muestras obtenidas de las distintas bases de datos se dividen en dos grupos: el grupo de diseño (*training set*) y el grupo de test (*test set*). Los clasificadores se entrenan con el conjunto de muestras del grupo de diseño, mientras que el error de clasificación se estima con el conjunto de test. El proceso se repite un determinado número de veces y el error de clasificación final se obtiene calculando la media de los errores conseguidos en estas repeticiones.

Las técnicas de clasificación pueden dividirse en dos categorías dependiendo de si utilizan los contornos de la prosodia (*prosody contours*), tales como las características de ésta de las ventanas de voz, o de si utilizan las estadísticas de estos contornos (*statistics of prosody contours*) como la media o la varianza.

Las técnicas de clasificación que utilizan el contorno de la prosodia estudian la información temporal de la voz y del discurso, mientras que aquellas que utilizan las estadísticas de este contorno se dividen entre las técnicas que estiman la función de densidad de probabilidad (*pdf*) y las que discriminan entre estados emocionales sin estimar la distribución de las características de la voz ya obtenidas.

En este apartado se describirá en qué consiste y cómo se aplican algunos de los métodos que pertenecen al primer grupo.

Técnica basada en redes de neuronas artificiales (*artificial neural networks, ANNs*)

En este clasificador, las características de las ventanas de voz se utilizan como entrada a estas redes de neuronas artificiales.

En este algoritmo cada muestra se divide en Q intervalos, los cuales contienen K elementos cada uno. El número de intervalos depende de la longitud de la muestra, mientras que el número de elementos se mantiene siempre constante.

Si denotamos u_ξ a una muestra obtenida de la base de datos y $x_{\xi q}$ a uno de los intervalos de la muestra, a una de las ventanas, utilizando el modelo oculto de Markov (*Hidden Markov Model, HMM*) $x_{\xi q}$ se clasifica automáticamente en uno de los grupos de fonemas: fricativos, vocales, semivocales, etc.

En cada uno de los elementos del intervalo $x_{\xi q}$, se extrae un número de características relacionadas con el estado emocional del discurso, este número se expresará con la letra D. De esta forma obtenemos una matriz KxD que, tras convertirla en un vector de $K \cdot D$ elementos, podemos introducir en el algoritmo ANN.

Este algoritmo se entrena en un estado emocional determinado, Ω_c, de un grupo de fonemas determinado, θ_λ. La salida del ANN indica la probabilidad condicionada de $x_{\xi q}$ dado el estado emocional Ω_c y el grupo de fonemas θ_λ.

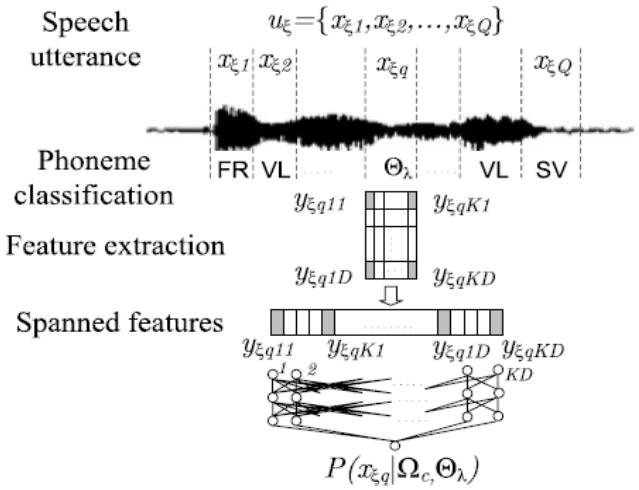

Figura 1. Algoritmo de clasificación de emociones que utiliza HMMs para la clasificación de fonemas y ANNs para la de emociones. Figura tomada de [3].

La probabilidad de una muestra, dado el estado emocional, se calcula de la siguiente manera:

$$P(u_\xi|\Omega_c) = \sum_{q=1}^{Q} \sum_{\lambda=1}^{\Lambda} P(x_{\xi q}|\Omega_c, \Theta_\lambda) P(\Theta_\lambda)$$

Modelo oculto multicanal de Markov (*multi-channel hidden Markov model, HMM*)

Si consideramos una secuencia de canales HMM simples, s_i donde i=1,2,…,V, un sistema de clasificación se puede definir, en cualquier momento, en uno de los V estados anteriores. Estos estados se corresponden con fonemas, es decir, el número de estados se corresponde al número de fonemas esperados en las muestras.

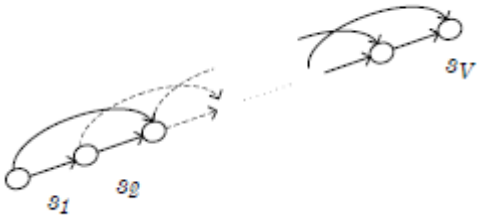

Figura 2. HMM de canal simple (single-channel HMM). Figura tomada de [3].

Como se puede observar en la figura anterior, la estructura de un HMM va siempre de izquierda a derecha puesto que los fonemas, dentro del discurso, siguen una secuencia de izquierda a derecha.

Si consideramos un sistema HMM de canales simples, la probabilidad de la muestra u_ξ dado el modelo λ para cada emoción, $P(u_\xi|\lambda)$, se calcula con la siguiente ecuación, en la que N es el número de estados HMM y $\alpha_T(i)$ es la variable terminal hacia adelante (*terminal forward variable*) obtenida a partir del algoritmo hacia adelante:

$$P(u_\xi|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

El algoritmo HMM multicanal, consiste en un conjunto de estados s_{cv} donde c=1,2,…,C y v=1,2,…,V. En estos estados, las variaciones de la componente c forman un disco. Dentro de cada disco, las transiciones de izquierda a derecha para pasar de un estado emocional a otro están permitidas. También están permitidas las transiciones de izquierda a derecha entre estados emocionales de un disco y el posterior.

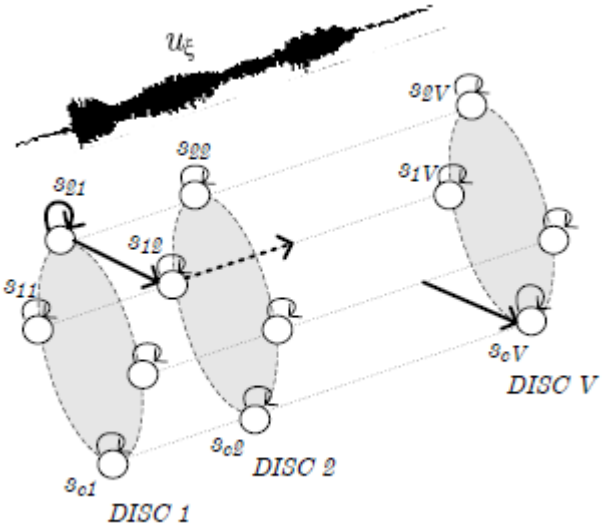

Figura 3. HMM multicanal (multi-channel HMM). Figura tomada de [3].

La fase de entrenamiento de este algoritmo se divide en dos pasos. En el primer paso se entrena cada uno de los canales simples para un estado emocional. El segundo paso combina los canales simples que dependen de las emociones para obtener un HMM multicanal.

Para clasificar una muestra, se calcula la probabilidad de la misma dado el estado emocional, Ω_c, como la relación del número de saltos entre estados, s_{cv} donde v=1,2,…,V, frente al número total de transiciones.

Mezcla de HMMs (*mixture of HMMs*)

Este último algoritmo también está compuesto por dos fases de entrenamiento. En la primera fase, se utiliza un algoritmo de clustering sin supervisar para obtener M clusters en el espacio de características de los datos de entrenamiento. Para ello se asume que la información de cada cluster está gobernada por un único HMM.

Sean $\Gamma^{(l)} = \{\gamma_1^{(l)}, \dots, \gamma_m^{(l)}, \dots, \gamma_M^{(l)}\}$ los clusters de la iteración número l del algoritmo de clustering, $\Delta^{(l)} = \{\delta_1^{(l)}, \dots, \delta_m^{(l)}, \dots, \delta_M^{(l)}\}$ los parámetros del HMM del anterior conjunto de clusters y $P\left(u_\xi \middle| \delta_m^{(l)}\right)$ la probabilidad de u_ξ dado el cluster con los parámetro del HMM $\delta_m^{(l)}$, el logaritmo de la verosimilitud de todas las muestras durante la iteración número l se calcula de la siguiente forma:

$$P^{(l)} = \sum_{m=1}^{M} \sum_{u_\xi \in \gamma_m^{(l)}} \log(P\left(u_\xi \middle| \delta_m^{(l)}\right))$$

En la segunda etapa, las muestras que han sido clasificadas en un cluster γ_m se utilizan para entrenar un número C de HMMs donde cada HMM se corresponde con un estado emocional.

Para clasificar una muestra del conjunto de test en uno de los diferentes estados emocionales se utiliza un clasificador de Bayes:

$$P(\Omega_c|u_\xi) = \sum_{m=1}^{M} P(\Omega_c, \delta_m|u_\xi) = \sum_{m=1}^{M} P(u_\xi|\Omega_c, \delta_m)P(\delta_m|\Omega_c)P(\Omega_c)$$

En este clasificador, $P(\delta_m|\Omega_c)$ es la relación entre las muestras que se han asignado al cluster γ_m y que pertenecen a Ω_c y el número de muestras del conjunto de entrenamiento, mientras que $P(u_\xi|\Omega_c, \delta_m)$ es la salida del HMM que fue entrenado en el estado emocional Ω_c y $P(\Omega_c)$es la probabilidad de cada estado emocional.

Support Vector Machine o máquinas kernel (*SVM*)

SVM es una técnica basada en la clasificación binaria, en la que únicamente tenemos dos hipótesis, aunque puede utilizarse para clasificaciones en las que se utilicen más de dos hipótesis.

SVM realiza la clasificación con un hiperplano (*hyperplane*) como superficie de decisión en el que $\boldsymbol{y}$ es el vector de medidas, $\boldsymbol{w}$ el vector del gradiente perpendicular a este hiperplano y b el offset del hiperplano desde el origen.

$$g(y) = \boldsymbol{w}^T \boldsymbol{y} + b$$

El hiperplano óptimo de separación (*optimal separation hyperplane, OSH*), el cual es uno de los criterios críticos para la clasificación, se define como un hiperplano que puede separar muestras o datos entre los que existen diferencias obvias. Para aquellas muestras que no pueden ser separadas por una línea recta, se introduce el parámetro regulatorio C. Este parámetro elegirá un hiperplano que permita una mala clasificación de los datos mientras máximiza el margen de la superficie de decisión.

El hiperplano se crea a partir de las muestras que pertenecen al conjunto de entrenamiento (o diseño) y a las que se conoce como vector de soporte (*support vector*). Cuando el número de este vector de soporte sea el más pequeño posible, el hiperplano estará situado en el punto óptimo.

Cuando los datos de los que se disponen en la práctica no son separables a partir de un hiperplano, es necesario utilizar funciones de kernel (*kernel functions*). El algoritmo del kernel transforma los puntos de entrada en puntos de grandes dimensiones antes de la clasificación. Las funciones más utilizadas son las de kernel lineal (*linear kernel*) y RBF kernel.

A la función del kernel lineal se la conoce como el producto escalar (*dot product*) y sus valores de salida son combinación lineal de todos los support vectors.

$$k(x_i, x_j) = x_i \cdot x_j$$

La función del kernel RBF cuya ecuación se muestra a continuación, proporciona una buena generalización así como buenos resultados al resolver este tipo de problemas.

$$k(x_i, x_j) = \exp\left(\gamma \left| x_i \cdot x_j \right|^2\right)$$

Para minimizar el error se puede elegir un margen mayor, γ, el cual representa la distancia entre el hiperplano y las muestras más cercanas a este.

Otros estudios

Para concluir este estado del arte sólo nos cabe mencionar que las herramientas anteriormente descritas se utilizan también para el reconocimiento de estados tales como la asfixia en el llanto de los niños [7]. Este tipo de estudios están centrados en reconocer en el tipo de llanto de un bebé, cuándo un niño está sufriendo algún tipo de enfermedad o ataque con el objetivo de reducir la mortalidad infantil por causas como la falta de oxígeno.

Otros estudios se basan en el reconocimiento de emociones en presencia de ruido. La mayor parte de las investigaciones realizadas sobre el reconocimiento de emociones se llevan a cabo utilizando bases de datos en las que las muestras de voz han sido grabadas en un estudio y, por tanto, en ausencia de ruidos tales como pasos en una habitación o voces de otras personas.

En este tipo de estudios se suelen utilizar bases de datos como la denominada SUSAS, que ha sido grabada en condiciones espontáneas y en la que las muestras se encuentran parcialmente ocultas por el ruido ambiental, o bases de datos grabadas en estudios controlados a las que se les ha añadido ruido blanco posteriormente.

3. EXTRACCIÓN DE LAS CARACTERÍSTICAS ACÚSTICAS PARA RECONOCIMIENTO DE EMOCIONES

Tal como se ha mencionado en la introducción de este proyecto, el objetivo principal del mismo se centra en la búsqueda y comparación de los resultados obtenidos a partir de diferentes características acústicas. De esta forma se pretendía que, al finalizar el mismo, pudiéramos ser capaces de ofrecer una serie de conclusiones acerca de la fiabilidad de estas características respecto al proceso de reconocimiento de emociones.

A lo largo de este capítulo vamos a realizar una descripción acerca de los diferentes parámetros utilizados en este proceso, teniendo en cuenta además que estos han formado parte de dos tipos distintos de experimentos (en trama y en segmento) que serán explicados más en profundidad en las siguientes secciones.

Mel Frequency Cepstral Coefficients (MFCC)

El objetivo de un análisis cepstral se basa en la extracción de la envolvente espectral de la señal de voz. Los coeficientes Mel-Frequency Cepstral representan el espectro de potencia a corto plazo (short-term) de un sonido. Estos coeficientes son el resultado de la transformada lineal de coseno del logaritmo de la potencia del espectro, en una escala Mel de frecuencia no lineal. Además, su objetivo se basa en deconvolucionar los efectos de la forma del tracto vocal y la excitación de las cuerdas vocales.

El primer paso para generar los coeficientes MFCC se basa en dividir la señal auditiva en tramas, generalmente de unos 20 milisegundos, de forma que la trama obtenida presente una serie de características que sean estadísticamente estacionarias. Este proceso se lleva a cabo mediante el enventanado de la muestra de audio. El tipo de ventana utilizado es una ventana Hamming puesto que este tipo de ventanas elimina los efectos de los bordes y proporciona mejores resultados [13].

Un nuevo conjunto de parámetros MFCC se genera a partir de cada una de las tramas obtenidas. Un esquema del proceso de obtención de estos coeficientes se muestra en la siguiente figura.

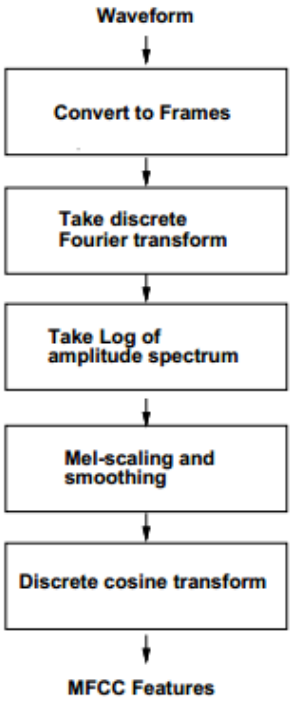

Figura 4. Proceso de obtención de los parámetros MFCC. Figura tomada de [13].

Una vez realizado el enventanado de la muestra de audio, se lleva a cabo la Transformada Discreta de Fourier en cada una de las tramas para obtener el espectro de frecuencia de la señal de voz y, posteriormente, se calcula su módulo al cuadrado, puesto que se considera que la fase del espectro no aporta información significativa.

El siguiente paso de este procedimiento consiste en suavizar el espectro y enfatizar las frecuencias significativas. Esto se consigue concentrando todas las componentes espectrales en una serie de canales de frecuencias (*frequency bins*), en nuestro caso 23 canales, en los que se promedian las componentes espectrales como se puede observar en la figura 5.

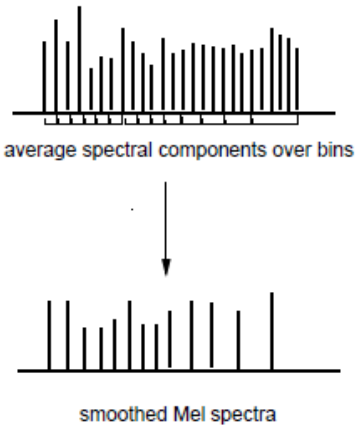

Figura 5. Proceso de agrupación de frecuencias y suavizamiento del espectro. Figura tomada de [13].

Estos canales no se encuentran equiespaciados puesto que se ha demostrado que, para el habla, las bajas frecuencias tienen mayor importancia que las altas y, debido a esto, estos canales siguen la escala de frecuencias Mel [14].

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

La escala de frecuencias de Mel se basa en el mapeo entre las frecuencias reales y la frecuencia percibida, puesto que el ser humano no percibe las frecuencias de forma lineal. Este mapeo es aproximadamente lineal por debajo de 1kHz y logarítmico por encima.

Para llevar a cabo la agrupación mencionada anteriormente, los parámetros previamente obtenidos se correlacionan con el banco de filtros triangulares de escala Mel.

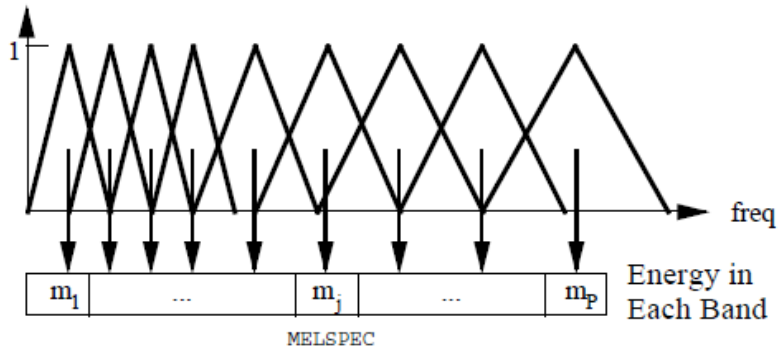

Figura 6. Banco de filtros de escala Mel. Figura tomada de [14].

A continuación, se calcula el logaritmo de las energías obtenidas a la salida de cada una de las bandas que componen el banco de filtros, con objeto de reducir su margen dinámico. Dichos parámetros se denominar log-energías en banda.

El último paso del proceso consiste en aplicar la Transformada Discreta de Coseno con el objetivo de decorrelar los parámetros obtenidos anteriormente y obtener los coeficientes MFCC finales (c_i).

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right)$$

En la ecuación anterior N es el número total de canales y m_j la log-energía en la banda j obtenida en el paso anterior. Puesto que los coeficientes cepstrales de mayor orden contienen poca información sobre el tracto vocal, habitualmente dichos coeficientes se descartan.

Frequency-Filtered Logarithmic Filter Bank Energies (FFLFBE)

Los parámetros FFLFBE fueron incluidos en nuestros experimentos, debido a que la comparación de los resultados obtenidos a partir de las log-energías en banda (parámetros FFLFBE) con los coeficientes MFCC (combinación lineal de estas log-energía en banda) mencionados en la sección anterior podría resultar interesante. [15]

La extracción de los coeficientes FFLFBE se lleva a cabo siguiendo un proceso muy similar al utilizado para la obtención de los parámetros MFCC, es decir, los primeros pasos de ambos procedimientos son los mismos. Sin embargo, en el caso de los FFLFBE, una vez llevada a cabo la agrupación de frecuencias siguiendo una escala Mel, la decorrelación de las log-energías en banda no se realiza a partir de la Transformada Coseno, sino utilizando un filtro paso banda, en nuestro caso un filtro FIR de segundo orden $H(z) = z - z^{-1}$.

Este filtro está diseñado para eliminar las componentes de gran varianza, además de la media de las log-energías en banda, debido a que tiene un cero en el origen. La respuesta al impulso de dicho filtro es la siguiente:

$$h[n] = \delta[n+1] - \delta[n-1]$$

Como puede observarse, este filtro no es un filtro causal. Debido a esto, es necesario añadir en los dos extremos del vector de log-energías en banda dos valores que representan las componentes de baja frecuencia en los límites inferior y superior de la banda total de frecuencia.

De esta forma los coeficientes FFLFBE $\left(X_t^{lf}\right)$ de la trama t serían:

$$X_t^{lf}(0) = X_t^l(1)$$

$$X_t^{lf}(m) = X_t^l(m+1) - X_t^l(m-1), \qquad 0 < m \leq N_b - 2$$

$$X_t^{lf}(N_b - 1) = -X_t^l(N_b - 2)$$

Donde X_t^l son las log-energías en banda y N_b es el número de bandas totales.

Pitch o frecuencia fundamental

Como hemos descrito anteriormente, la frecuencia fundamental de la voz se define como la tasa de vibración de las cuerdas vocales y ésta contiene información acerca de las emociones. Esto es debido a que depende de la tensión de las cuerdas vocales además de la presión del aire procedente de la cavidad pulmonar. Una explicación más extensa sobre esta característica puede encontrarse en la sección *Estado del Arte*.

La elección de incluir esta característica en diferentes experimentos de este proyecto, se basó en el conocimiento de que este parámetro contiene información acerca de la prosodia, de forma que podría ayudarnos a conseguir mejores resultados.

Para obtener la frecuencia fundamental de cada una de las tramas, primero enventanamos la señal de audio, tal como se ha explicado anteriormente, obteniendo de esta manera tramas de 20 milisegundos. Una vez obtenidas estas tramas, utilizamos la función de Matlab $fxrapt$ para calcular la frecuencia fundamental de cada una de ellas. En el caso de que la trama sea sorda, se considera que su frecuencia fundamental es 0.

Versiones Segmentales: Integración Temporal de Características (Temporal Feature Integration)

Tal como se ha mencionado en la introducción de este capítulo, en este proyecto hemos llevado a cabo dos versiones de los diferentes experimentos realizados, una versión en trama y una versión en segmento, y en ambas versiones hemos utilizado los parámetros previamente comentados.

La versión en trama de los experimentos se corresponde con el cálculo de los parámetros tal y como se ha explicado en las secciones anteriores de este capítulo. Un conjunto de parámetros se obtiene para todas y cada una de las tramas en las que se ha dividido la señal de audio.

Los coeficientes fueron obtenidos en tramas de 20 segundos con un solapamiento de 10 segundos entre ellas. En concreto, el número total de parámetros por trama fue: 12 coeficientes en el caso de los MFCC, 23 en el de los FFLFBE, 1 coeficiente que contenía la energía de la trama y 1 en el caso del pitch.

En cuanto a la versión segmental de estos experimentos, primero calculamos su versión en trama y a continuación, aplicamos sobre dichos parámetros a nivel de trama diversas técnicas de integración temporal de características en segmentos de 60, 100, 300, 600 y 1000 milisegundos.

En las siguientes subsecciones se describen los métodos de integración de características utilizados, los cuales hemos dividido entre métodos basados en estadísticos y métodos basados en bancos de filtros.

Métodos basados en estadísticos

Media (Mean)

Estadísticamente, la media se refiere a la medida de la tendencia central de una distribución de probabilidad o de una variable aleatoria caracterizada por dicha distribución.

En el caso de una distribución de probabilidad discreta, la media de una variable aleatoria X se corresponde con la suma de cada valor posible de la misma, x, multiplicado por la probabilidad de ese valor, $P(x)$, donde $\mu = \sum x \cdot P(x)$. En el caso de una distribución de probabilidad continua, se aplica una fórmula análoga para calcular la media, $\mu = \int_{-\infty}^{\infty} x \cdot f(x)\ dx$, donde x es la variable aleatoria y $f(x)$ es la distribución de probabilidad continua.

En el caso que nos ocupa, la fórmula utilizada para calcular la media se obtiene teniendo en cuenta que contamos con una distribución de probabilidad discreta en la que todos sus valores son equiprobables y, por tanto, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Desviación Típica (Standard deviation)

La desviación típica representa cuánta variación, o dispersión, existe respecto a la media. Cuando este parámetro es muy pequeño nos indica que los datos tienden a tener un valor muy próximo a la media, mientras que una desviación típica alta muestra que los datos toman valores muy dispersos.

La desviación típica se define como la raíz cuadrada de la varianza (la varianza de una variable aleatoria, al igual que la desviación típica, es una medida de dispersión definida como la esperanza al cuadrado de la desviación de dicha variable respecto a su media), siendo su ecuación mostrada a continuación.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Donde σ es la desviación típica, μ la media y N el número total de muestras.

Skewness

El parámetro estadístico conocido como skewness, también llamado asimetría, es una medida del grado de asimetría que presenta una distribución de probabilidad.

Con respecto a este parámetro, se puede afirmar que si la cola izquierda de la distribución, la parte final izquierda de la misma, es más pronunciada que la de la derecha se establece que esta función tiene una skewness negativa, mientras que si ocurre lo contrario se trata de una skewness positiva. En el caso en el cual los dos extremos de la distribución son iguales decimos que la distribución tiene cero skewness.

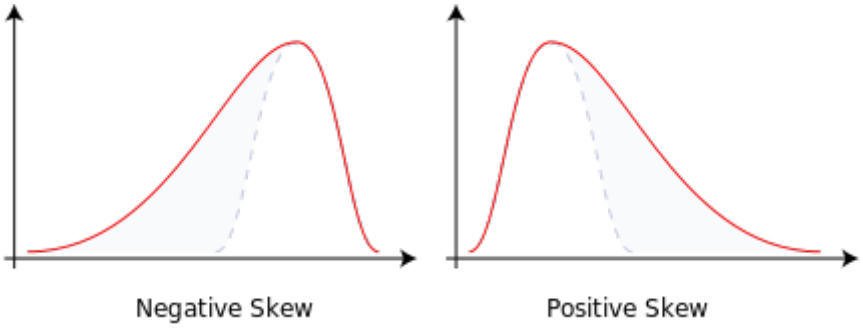

Figura 7. Skewness negativa y positiva respectivamente.

La ecuación de este parámetro es la siguiente, $\gamma_1 = \frac{\mu_3}{\mu_2^{2/3}}$, donde μ_i es el iésimo momento central $\mu_i = \langle (x - \langle x \rangle)^i \rangle$ y, en particular, μ_2 es la varianza.

Curtosis (Kurtosis)

El parámetro estadístico denominado curtosis mide el grado de apuntamiento de una distribución o de una variable aleatoria de valores reales.

Este parámetro se define como la forma normalizada del cuarto momento central de una distribución de probabilidad y es un descriptor de la forma de dicha distribución.

Aunque existen diferentes formas de medir la curtosis y, por tanto, diferentes ecuaciones son utilizadas para su cálculo, el más conocido es $\beta_2 = \frac{\mu_4}{\mu_2^2}$, donde μ_i es el iésimo momento central y, en particular, μ_2 es la varianza.

Método basado en banco de filtros

Este método es un tipo de integración temporal de características, proceso en el cual se combinan todos los vectores de características de un segmento (secuencia de varias tramas) de forma que se genere un nuevo vector que contenga toda la información relevante de dicho segmento.

Para llevar a cabo la adquisición de dichos parámetros o características seguimos el proceso indicado en [16]. Primero, establecimos un único vector, z_k, en el cual se englobara la secuencia de parámetros a nivel de trama obtenidos en un primer momento. En nuestro caso, estos parámetros se refieren a los coeficientes MFCC y FFLFBE, mencionados en un apartado anterior, puesto que llevamos a cabo sendos experimentos con ellos.

$z_k = f(x_{k \cdot h_{s_x}}, x_{k \cdot h_{s_x}+1}, \dots, x_{k \cdot h_{s_x}+f_{s_x}-1})$, donde $k = 0, 1, \dots, K-1$, f_{s_x} es el tamaño de la trama, h_{s_x} es el tamaño del solapamiento y x los coeficientes MFCC o FFLFBE.

Con el fin de llevar a cabo la integración temporal de características utilizando este método, se estimará primero el espectro de potencia de los parámetros contenidos en el segmento k-ésimo calculando para ello el periodograma de cada una de sus componentes. Teniendo esto en cuenta, podemos establecer que $z_k^{(i)}$es el periodograma de dimensión D_z del iésimo coeficiente MFCC o FFLFBE, donde $D_z = \frac{f_{S_x}}{2} + 1$.

Una vez obtenido el espectro de potencia de dichos coeficientes, se agrupará la energía en diferentes bandas de frecuencia utilizando para ello un banco de filtros predefinido, $\boldsymbol{W}$, y se obtendrá el vector de características, $\check{\mathrm{z}}_k^{(i)}$. Éste tendrá el mismo número de componentes que filtros hay en el banco y se utilizará como entrada al proceso de reconocimiento de emociones.

$$\check{\mathrm{z}}_k^{(i)} = \boldsymbol{W}^T z_k^{(i)}$$

En nuestro caso, dicho banco de filtros engloba cuatro bandas de frecuencia:

1. $0\ Hz$ (valor DC).
2. $1 - 2\ Hz$ (beat rates).
3. $3 - 15\ Hz$ (energía de modulación).
4. $20 - \frac{S_x}{2}\ Hz$ (perceptual roughness), donde s_x es la frecuencia de muestreo.

En la siguiente figura se muestra un resumen del proceso anteriormente descrito.

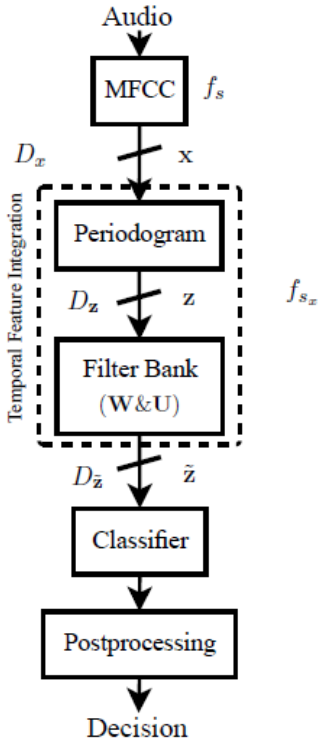

Figura 8. Proceso de la integración temporal de características. Figura tomada de [16].

4. SISTEMA DE RECONOCIMIENTO DE EMOCIONES Y BASES DE DATOS

Este cuarto capítulo se encuentra dividido en dos secciones. En la primera sección, vamos a llevar a cabo una descripción del sistema de reconocimiento de emociones utilizado en este proyecto, mientras que en la segunda se describirán las características de las dos bases de datos a partir de las cuales hemos obtenido nuestras muestras de audio.

En la siguiente figura se puede observar el diagrama de bloques del sistema de reconocimiento de emociones utilizado.

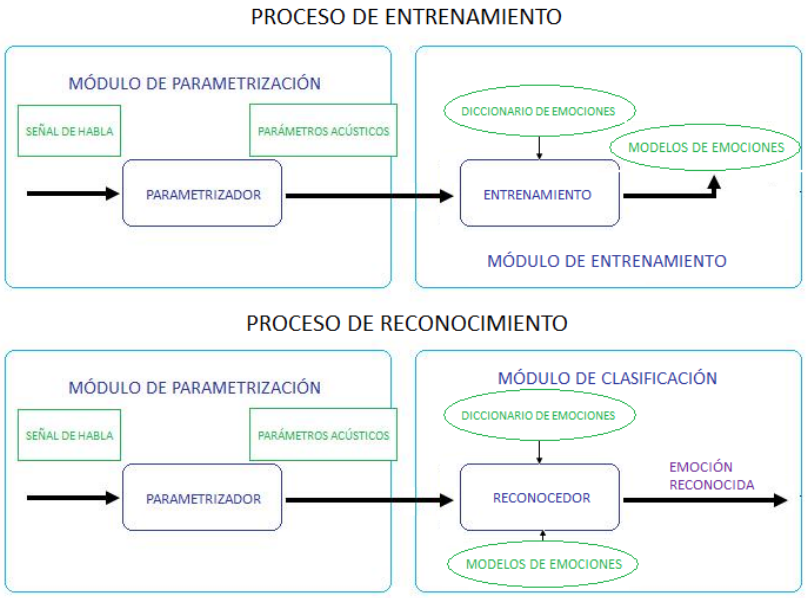

Figura 9. Diagrama de bloques del sistema de reconocimiento de emociones.

Sistema de Reconocimiento de Emociones

Una vez obtenidos los parámetros mencionados en el capítulo anterior, procedimos a realizar el reconocimiento de emociones propiamente dicho. Para ello, este proceso se dividió en dos partes: entrenamiento del sistema y reconocimiento de emociones.

Ambos procesos se llevan a cabo utilizando una herramienta denominada HTK [14] cuya principal función es la de construir sistemas para el procesamiento del lenguaje basados en modelos ocultos de Markov (Hidden Markov Models, HMM). Como se ha mencionado anteriormente en la sección denominada *Modelo oculto multicanal de Markov* del capítulo *Estado del Arte*, un HMM es un modelo estadístico en el que se asume que el sistema que se va

modelar es un proceso de Markov cuyos parámetros no se conocen y cuyo objetivo es determinar estos parámetros.

En primer lugar, HTK considera que cada palabra, o cada emoción en nuestro caso, puede ser representada por una secuencia de vectores O donde o_t es el vector observado en el instante$t$.

$$O = o_1, o_2, \dots o_T$$

En el caso de llevar a cabo el reconocimiento de palabras aisladas, la solución a este problema se obtiene a través del siguiente cálculo:

$$\arg\max_i \{P(w_i|O)\}$$

En la ecuación anterior w_i es la iésima palabra del vocabulario. Como se puede observar no es posible calcular dicha probabilidad directamente pero aplicando el teorema de Bayes obtenemos:

$$P(w_i|O) = \frac{P(O|w_i)\,P(w_i)}{P(O)}$$

Si las probabilidades a priori, $P(w_i)$, son conocidas podemos observar que el clasificador anterior sólo depende de la probabilidad condicional $P(w_i|O)$. Esta probabilidad condicional se estima a partir de la obtención de los parámetros de un modelo de Markov. Un modelo de Markov es una máquina de estados finita que cambia de estado en cada instante de tiempo. Cada instante de tiempo t en que se entra en un nuevo estado j, se genera un vector o_t a partir de la densidad de probabilidad $b_j(o_t)$. Además la probabilidad de pasar de un estado i a un estado j está representada por la probabilidad discreta a_{ij}.

De esta forma la probabilidad conjunta que representa que el vector O es generado por el modelo de Markov M siguiendo una secuencia de estados X se obtiene de la siguiente manera:

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3) \dots$$

En la práctica la secuencia de estados X no es conocida, por lo tanto, nuestro modelo se puede representar a partir de un Modelo Oculto de Markov (*HMM*).

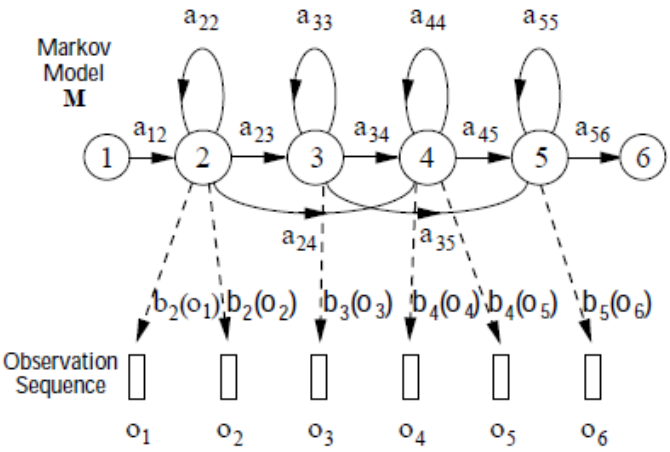

Figura 10. Modelo de Markov. Figura tomada de [14].

Puesto que la secuencia de estados X es desconocida, la probabilidad condicional se obtiene teniendo en cuenta todas las posibles secuencias de estados $X = x(1), x(2), x(3), \dots, x(T)$.

$$P(O|M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

Donde $x(0)$ es la entrada del modelo y $x(T+1)$ la salida del mismo. Estos dos estados no generan un vector de observación como se puede comprobar en la figura previa.

Otra forma de calcular esta probabilidad condicional es tener en cuenta únicamente la secuencia de estados más probable:

$$P(O|M) = \max_{X} \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right\}$$

Teniendo en cuenta que un conjunto de modelos de Markov M_i se corresponden con un conjunto de palabras o, en nuestro caso, de emociones w_i, podemos asumir que:

$$P(O|w_i) = P(O|M_i)$$

Todo esto asume que los parámetros a_{ij} y $b_j(o_t)$ son conocidos para cada uno de los modelos M_i. En el caso que nos ocupa, los parámetros de un HMM pueden obtenerse a partir de un conjunto de muestras de entrenamiento que pertenezcan a dicho modelo y a partir de un algoritmo robusto y eficiente de re-estimación.

En la siguiente figura se muestra cómo se lleva a cabo el proceso de reconocimiento de emociones utilizando para ello un Modelo Oculto de Markov. Primero, se construye un HMM por cada una de las emociones que forman parte del conjunto que vamos a reconocer utilizando para ello las muestras de entrenamiento y, posteriormente, se calcula la probabilidad condicional de la emoción a reconocer para cada uno de los modelos anteriormente generados y la más probable es la que identificará el modelo de dicha emoción.

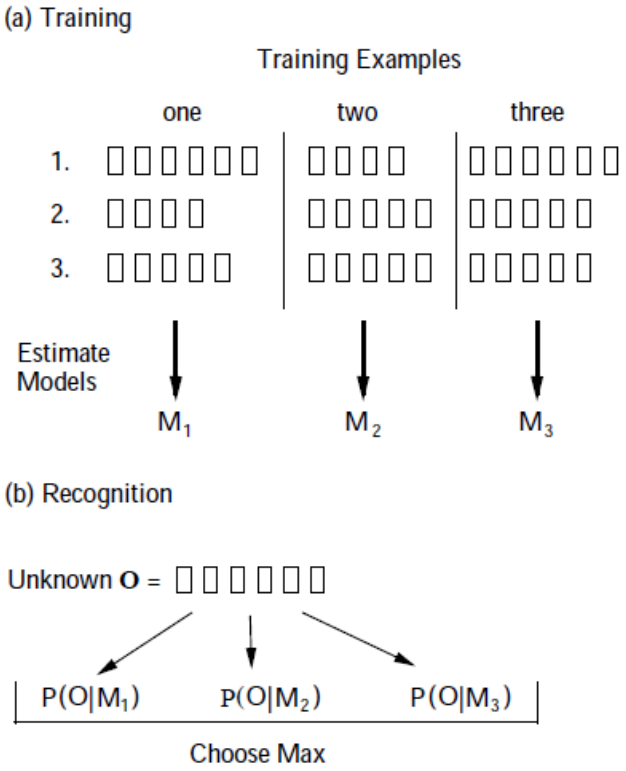

Figura 11. Reconocimiento de emociones utilizando HMMs. Figura tomada de [14].

A lo largo tanto de la fase de entrenamiento como de la de reconocimiento la distribución de salida, $b_j(o_t)$, se representa mediante una mezcla de Gaussianas (Gaussian Mixture Model, *GMM*). Las GMM se utilizan en el reconocimiento de habla como un modelo probabilístico genérico para densidades multivariables capaz de representar densidades arbitrarias.

En el caso que nos ocupa, cada vector de observación en el instante t, o_t, se divide en un número independiente de streams, o_{st}, donde S es el número total de streams. La fórmula utilizada para calcular esta probabilidad de salida es la siguiente:

$$\mathrm{b_j}(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s}$$

En la fórmula anterior M_s es el número de Gaussianas de la mezcla para cada trama s, c_{jsm} el peso de la emésima componente, γ_s el peso del stream y $\mathcal{N}(o; \mu, \Sigma)$ una Gaussiana multivariante con media el vector μ y matriz de covarianza Σ. De esta forma la fórmula de dicha Gaussiana es la siguiente:

$$\mathcal{N}(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)' \Sigma^{-1} (o-\mu)}$$, donde n es la dimensión de o.

Los parámetros de estas mezclas de Gaussianas se calculan utilizando para ello el algoritmo de Re-Estimación de Baum-Welch [17 – 21] y el algoritmo de Viterbi [22] además, en nuestro caso se utiliza un stream de peso 1.

Bases de Datos

Para llevar a cabo este proyecto, se había planteado en principio la utilización de una única base de datos de emociones sobre la cual realizar los distintos experimentos, siendo la base de datos EMO-DB elegida para ello. Sin embargo, a medida que estos experimentos fueron llevándose a cabo comenzamos a pensar que podría ser interesante realizar los mismos sobre una nueva base de datos y poder así comparar los resultados obtenidos. Fue por este motivo por el cual se incluyó la base de datos SAVEE en este proyecto.

Berlin Emotional Speech Database (*EMO-DB*)

La base de datos EMO-DB constituye una base de datos alemana en la que las muestras contenidas engloban 7 emociones diferentes que han sido simuladas por actores en un entorno de grabación controlado. Las emociones presentes son las siguientes: neutralidad, enfado, miedo, alegría, tristeza, asco y aburrimiento [10].

Para establecer esta base de datos diez actores, cinco hombres y cinco mujeres elegidos debido a su naturalidad y a la facilidad de llevar a cabo el reconocimiento de las emociones en sus actuaciones, produjeron diez muestras auditivas cada uno en las cuales simulaban las diferentes emociones propuestas. Estas diez muestras se recopilaron en alemán y se componen de cinco frases cortas y cinco largas que pueden ser utilizadas en la vida diaria y pueden aplicarse en todas y cada una de las diferentes emociones.

La grabación de estas frases, además de sus correspondientes electro-glotogramas (electro-*glottograms*), se llevó a cabo en la cámara anecoica de la Universidad Técnica de Berlín (*Technical University Berlin*), utilizando para ello un equipo de grabación de alta calidad y un laringógrafo portátil. Las muestras fueron tomadas con una frecuencia de muestreo de 48kHz y después submuestreadas a una frecuencia de 16kHz.

A lo largo del proceso de grabación, los actores se encontraban de pie en frente de un micrófono, de forma que les fuera posible gesticular si fuera necesario, con la única restricción de hablar en la dirección en la que se encontraba el micrófono a una distancia de 30cm.

El material completo se compone de unas 800 muestras que fueron evaluadas utilizando para ello un test de percepción centrado en la capacidad de reconocimiento de las diferentes emociones y la naturalidad de las mismas. Las muestras cuyos resultados respecto a la capacidad de reconocimiento fueron mayores del 80% y fueron, además, declaradas como naturales por

más del 60% de los encuestados en el test fueron etiquetadas fonéticamente. Esto supuso que unas 500 muestras de entre las 800 iniciales fueran descartadas.

Para garantizar la calidad emocional de las muestras así como su naturalidad, se llevó a cabo un test de percepción de las mismas de forma que los 20 sujetos que participaron en el mismo debían establecer a qué emoción correspondía cada una de las muestras.

A continuación se muestra una figura con las tasas de reconocimiento de las muestras en la que la línea que une diferentes emociones representa diferencias significativas entre ellas.

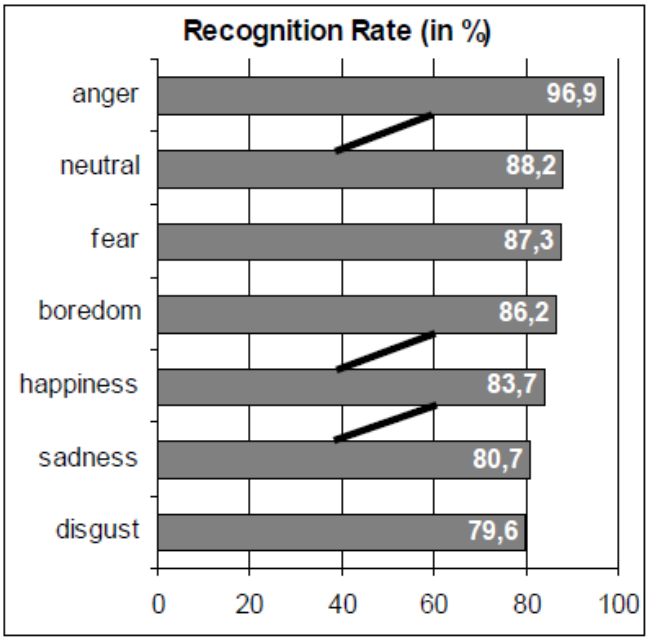

Figura 12. Tasa de reconocimiento en la base de datos EMO-DB. Figura tomada de [10].

Además de este, se llevaron a cabo dos tests más sobre los datos obtenidos. En uno de ellos se pidió a los participantes que puntuaran la intensidad de cada una de las emociones mostradas, mientras que en el otro se pidió que juzgaran la sílaba con mayor grado de acentuación de cada muestra.

Por último, para el etiquetado de las muestras, se crearon dos archivos en formato ASCII. El primero de estos archivos recoge la transcripción fonética de las mismas, mientras que el segundo contiene una segmentación en sílabas y marcadores de cuatro niveles de acentuación.

Surrey Audio-Visual Expressed Emotions (*SAVEE*)

La base de datos SAVEE fue creada con el fin de ser utilizada en el desarrollo de un sistema automático de reconocimiento de emociones multimodal. Está compuesta por 480 muestras en idioma inglés producidas por 4 actores, todos ellos hombres y estudiantes de posgrado además de investigadores en la Universidad de Surrey, en las que se representan 7 emociones diferentes (enfado, asco, miedo, alegría, tristeza, sorpresa y neutralidad) [11].

El material textual está compuesto por 15 frases TIMIT (TIMIT es una colección de transcripciones léxicas y fonéticas de discursos de hablantes ingleses americanos de diferentes sexos y dialectos en el que cada elemento descrito ha sido alineado temporalmente [12]) para cada emoción: 3 comunes, 2 específicas de la emoción en cuestión y 10 genéricas que son distintas para cada emoción, resultando en un total de 120 muestras por cada uno de los actores. Al tratarse de un recurso audiovisual, incluye, además de frases fonéticamente equilibradas (*phonetically-balanced*), 60 marcadores faciales de las caras de dichos actores.

Todos los datos fueron grabados en un laboratorio de medios visuales utilizando para ello un equipo de alta calidad audiovisual además de ser, posteriormente, procesados y etiquetados. Las transcripciones de las frases que debían interpretar los actores, así como un vídeo y 3 imágenes que pretendían provocar la emoción deseada, se mostraban en un monitor delante de ellos durante las grabaciones. Con el fin de extraer las características de las expresiones faciales, la cara de los actores era pintada con 60 marcadores que se situaban en la frente, cejas, mejillas, labios y mandíbula como se muestra en la siguiente figura.

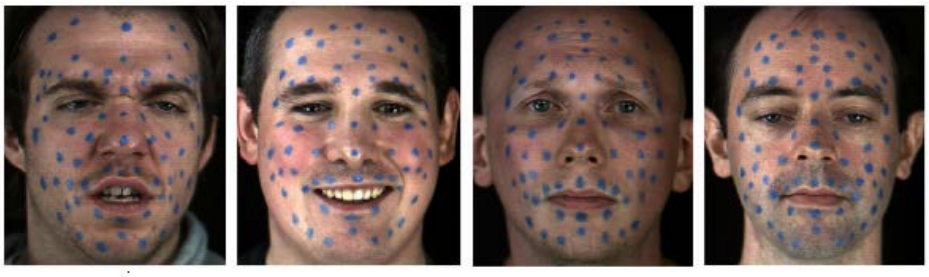

Figura 13. Marcadores azules utilizados para la grabación de las expresiones faciales en las distintas emociones. Figura tomada de [11].

La frecuencia de muestreo es de 44.1 kHz para audio y 60 fps para video.

Las muestras de audio fueron etiquetadas en dos pasos semi-automáticos. Primero, se utilizó el software HTK para etiquetar automáticamente las grabaciones y después, el software Speech Filling System (SFS) fue utilizado para corregir manualmente los errores de etiquetación. Con respecto a las muestras visuales, los marcadores fueron etiquetados manualmente en la primera trama de cada vídeo y, posteriormente, se procedió al seguimiento de los mismos en el resto de tramas.

Por último, el control de la calidad de estas actuaciones fue llevado a cabo por 10 sujetos de los cuales 5 eran hablantes nativos de inglés y el resto habían vivido en el Reino Unido más de un año, utilizando para ello bancos de pruebas con muestras únicamente auditivas, visuales y audiovisuales. Los resultados de estos tests, los cuales muestran que hay una mayor precisión en la clasificación de los datos visuales con respecto a los auditivos, fueron almacenados como resultados de referencia (*baseline*) de sistemas dependientes del hablante e independientes del mismo.

5. RESULTADOS EXPERIMENTALES

En este capítulo vamos a mostrar el conjunto de resultados obtenidos a partir de los experimentos que han sido llevados a cabo a lo largo de la realización de este proyecto. Los datos utilizados para ello han sido los almacenados en las dos bases de datos mencionadas anteriormente.

Así mismo, describiremos a grandes rasgos los procedimientos utilizados para ello y compararemos entre sí los resultados proporcionados por ambas bases de datos, para poder obtener unas conclusiones más sólidas.

Sistema de referencia

En el proceso general utilizado en este proyecto, el primer paso consiste en la parametrización completa de la base de datos sobre la cual se va a realizar el experimento. Con este fin, utilizamos la herramienta HTK para llevar a cabo gran parte de los procedimientos explicados en la sección *Extracción de las características acústicas para el reconocimiento de emociones*.

La obtención de características a partir de los ficheros contenidos en una de nuestras bases de datos se lleva a cabo con la ayuda del comando HCopy de HTK. Este comando se utiliza para copiar uno o más archivos fuente en un archivo de salida, sin embargo, si al ejecutar este comando se especifica un archivo de configuración es posible parametrizar todos los archivos de entrada.

A continuación se muestran algunos de los parámetros utilizados en la extracción de estas características. Los valores de los mismos pueden ser modificados en distintos experimentos, por lo que, en caso de que ocurriese, sería mencionado en las secciones correspondientes.

TARGETKIND = MFCC_E_D_A_Z

SOURCEFORMAT = WAV

SOURCERATE = 625

SOURCEKIND = WAVEFORM

TARGETFORMAT = HTK

TARGETRATE = 100000.0

WINDOWSIZE = 250000.0

USEHAMMING = T

ENORMALISE = T

ZMEANSOURCE = T

PREEMCOEF = 0.97

NUMCHANS = 40

NUMCEPS = 12

Los parámetros SOURCEFORMAT y SOURCEKIND indican que los que archivos de entrada utilizados en el reconocimiento de emociones van a ser ficheros de audio (fomas de onda) en formato WAV, mientras que SOURCERATE establece el periodo de muestreo del audio en 62,5 µsegundos.

TARGETKIND es una característica de configuración que determina el tipo de parámetros que se van a extraer de los archivos de entrada. En este caso (MFCC_E_D_A_Z), se trata de parámetros MFCC con normalización de la media (Z), con deltas (D) que se corresponden con las primeras derivadas de estos parámetros, aceleraciones (A) las cuales son las segundas derivadas de los mismos y la log-energía (E). NUMCEPS indica el número de coeficientes Cepstrales, el número de parámetros MFCC, que se van a calcular y que son 12 en nuestro caso, mientras que NUMCHANS indica el número de filtros del banco de filtros. TARGETFORMAT establece que los ficheros de salida son, en este caso, ficheros HTK y TARGETRATE la tasa de extracción de los parámetros que se establece en 10 milisegundos.

El parámetro USEHAMMING, el cual se encuentra fijado a “true” (T), indica que para el enventanado de las muestras de audio se va a utilizar una ventana de Hamming de 25 milisegundos, esto último establecido a partir del parámetro WINDOWSIZE. ZMEANSOURCE implica que las muestras van a tener una media nula antes de ser analizadas y PREEMCOEF fija el coeficiente del filtro de preénfasis. Por último, ENORMALISE cuyo valor es “true” (T) indica la normalización de la log-energía.

Una vez finalizada la parametrización de los ficheros se lleva a cabo el entrenamiento de los modelos para las distintas clases de emociones consideradas.

Los experimentos se realizan utilizando el paradigma denominado “Leave-One-Subject-Out” (LOSO), tal y como se explica a continuación. Los datos existentes en cada una de las bases de datos se dividen en diferentes grupos (10 grupos para la base de datos EMODB y 4 para la base de datos SAVEE). Cada uno de estos grupos incluye las muestras de voz producidas por un único individuo, de forma que en ningún subexperimento se utilice al mismo interlocutor para realizar el entrenamiento de nuestro reconocedor y el reconocimiento de emociones y gracias a esto, el proceso es independiente del hablante.

A partir de 9 de estos grupos para la base de datos EMODB y 3 para SAVEE, se generan un conjunto de modelos de Markov (un modelo de Markov por cada emoción) y se lleva a cabo el entrenamiento de los mismos a partir de las diferentes muestras de la base de datos y su correspondiente conjunto de etiquetas. En nuestro caso, se realizan mezclas de 1, 2, 4, 8, 16, 32,

64, 128 y 256 Gaussianas calculando los parámetros de las mismas utilizando el algoritmo de Re-Estimación Baum-Welch.

Una vez los distintos modelos han sido creados, se llevan a cabo los tests de reconocimiento de emociones para cada uno de ellos sobre el grupo restante, teniendo siempre en cuenta que en las muestras de test no están incluidas aquellas producidas por los interlocutores que generaron las muestras que forman parte del proceso de entrenamiento de estos modelos. El proceso de reconocimiento se realiza utilizando el algoritmo de Viterbi.

A partir de este procedimiento se genera un fichero de resultados por cada uno de los tipos de modelos (1, 2, 4, 8, 16, 32, 64, 128 y 256 Gaussianas) y cada uno de los grupos existentes. Por último, se calculan las tasas de reconocimiento finales producidos por cada uno de estos modelos de mezclas de gaussianas promediando los resultados parciales obtenidos en cada uno de los grupos involucrados en este proceso.

De esta forma se obtienen los datos que se presentan a continuación.

Resultados de la base de datos EMODB

Experimentos con los parámetros MFCC

En esta sección se van a mostrar los resultados obtenidos a partir de los experimentos llevados a cabo sobre los parámetros MFCC.

Tal como se ha explicado en el tercer capítulo de este proyecto, hemos desarrollado dos tipos de experimentos para cada uno de nuestros conjuntos de parámetros, los cuales han sido clasificados de la siguiente manera: por trama y por segmento.

Por trama

Este primer apartado contiene los resultados de referencia o baseline obtenidos para este proyecto así como los resultados de los diversos experimentos en los que la señal acústica se ha dividido en un conjunto de tramas.

Como primer paso para comenzar a definir el valor de diversos parámetros, decidimos comparar los resultados de dos experimentos cuya única diferencia radicaba en el número de filtros del banco de filtros utilizado. Esta característica se ve reflejada en el parámetro de nombre NUMCHANS.

De esta forma, diseñamos un experimento en el que el número de filtros fuese 40 y otro en el que éste fuese 23. Este segundo número se corresponde con el número de bandas críticas del sistema auditivo humano. Lo que se pretendía con este experimento era discernir si con un

número menor de canales en el banco de filtros podíamos obtener aproximadamente los mismos resultados, es decir, si el tamaño del banco de filtros utilizado era un factor determinante para la precisión de nuestro reconocedor de emociones.

A continuación se muestran los resultados obtenidos con respecto a este caso.

	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas
MFCC_E _D_A _Z 40 filtros	74,77%	74,95%
MFCC_E_D_A _Z 23 filtros	74,21%	74,50%

Tabla 4. Resultados de los experimentos con distinto número de filtros a partir de los parámetros MFCC para la base de datos EMODB.

Al estudiar la tabla anterior se puede observar cómo se produce una pequeña mejora en los resultados de los experimentos llevados a cabo con 40 filtros frente a los obtenidos utilizando únicamente 23. Sin embargo, al no tratarse de una diferencia muy importante, decidimos utilizar en nuestros experimentos 23 bandas de forma que nuestro modelo tuviese un mayor parecido con el sistema auditivo humano.

Teniendo en cuenta los resultados previos, el número de canales empleados en el resto de experimentos llevados a cabo a lo largo de este proyecto, tanto en aquellos en los que fueron utilizados los parámetros MFCC como en los que lo fueron los parámetros FFLFBE, fue 23.

Una vez determinado el valor de esta característica, se llevaron a cabo una serie de experimentos que incluían distintas combinaciones de parámetros, con el fin de averiguar cómo afectan al reconocimiento de emociones.

Para explicar la nomenclatura utilizada en las diversas tablas utilizaremos el siguiente ejemplo: MFCC_E_D_A_Z. En éste, MFCC se refiere a los parámetros calculados, E implica el cálculo de la energía de la trama, D los deltas correspondientes a la primera derivada de los parámetros MFCC, A las aceleraciones correspondientes a la segunda derivada de los mismos y Z implica que los coeficientes se han normalizado respecto a la media (es decir que la media de los coeficientes es 0). Además el parámetro “pitch” que aparece en algunos de los experimentos se refiere a la inclusión dentro del vector de parámetros de la frecuencia fundamental de cada una de las tramas. Este cálculo de la frecuencia fundamental se llevó a cabo a partir de la función $fxrapt$ de Matlab.

En la siguiente tabla se muestran los resultados obtenidos.

	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas
MFCC_Z	59,44%	60,19%
MFCC_E_Z	64,86%	65,61%
MFCC_D_Z	68,41%	69,16%
MFCC_E_D _Z	73,27%	73,08%
MFCC_E_D_A _Z	74,21%	74,50%
MFCC_E_D_Z + pitch	67.10%	68.41%

Tabla 5. Resultados de los experimentos por trama a partir de los parámetros MFCC para la base de datos EMODB.

Como se puede observar en la tabla anterior, tanto la presencia de la energía (E) como la de los deltas (D) supone una mejora notable en el porcentaje de aciertos, con respecto a los resultados obtenidos en ausencia de estos. La combinación de ambos parámetros produce a su vez un aumento en la precisión del reconocimiento de emociones.

Con respecto a las aceleraciones, a pesar de dar lugar a mejores resultados no se trata de un conjunto de parámetros tan críticos como son las deltas, mientras que al incluir el pitch de cada una de las tramas nuestros resultados empeoran.

De esta forma podemos concluir que para obtener mejores resultados en el reconocimiento de emociones a partir de señales acústicas utilizando los parámetros MFCC, la energía de la trama así como sus deltas deben estar incluidas como características en dicho proceso.

Por segmento

En este segundo apartado dividimos las tramas que contienen las características, obtenidas siguiendo el procedimiento explicado en la sección *Versiones Segmentales: Integración Temporal de Características (Temporal Feature Integration)* del apartado *Extracción de las características acústicas para reconocimiento de emociones*, en segmentos de 100, 300, 600 o 1000 milisegundos y realizamos una serie de modificaciones a los parámetros MFCC previamente calculados. Una vez concluida dicha modificación, se procede con el entrenamiento del reconocedor de emociones y el propio reconocimiento de las mismas siguiendo el mismo proceso que para los experimentos realizados por trama.

De esta forma, llevamos a cabo dos tipos de experimentos: basados en estadísticos y basados en bancos de filtros.

Experimentos basados en estadísticos

En este conjunto de experimentos, una vez definidos los segmentos a partir de los cuales van ser calculados los diversos parámetros, llevamos a cabo el cálculo de la media, desviación típica, skewnes y/o curtosis utilizando para ello una serie de funciones previamente definidas en Matlab. Por último, utilizamos el resultado de estos cálculos como características de entrada tanto para el entrenamiento como para el reconocimiento de las emociones.

A continuación se muestra una tabla que recoge los resultados de estos experimentos.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
		Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas
MFCC_Z media y desviación típica	60 ms	63,55%	67,10%	65,79%
	100 ms	66,92%	65,23%	59,81%
	300 ms	58,32%	55,14%	50,65%
	600 ms	50,65%	47,10%	41,12%
	1000 ms	51,21%	45,23%	36,64%
MFCC_Z media, desviación típica y skewness	60 ms	63,93%	66,36%	62,43%
	100 ms	64,86%	63,74%	57,38%
	300 ms	61,12%	55,14%	50.47%
	600 ms	57,01%	49,72%	44,67%
	1000 ms	51,78%	44,30%	38,50%
MFCC_Z media, desviación típica, curtosis y skewness	60 ms	62,06%	60,00%	58,69%
	100 ms	56,82%	56,45%	51,40%
	300 ms	56,26%	51,78%	45,98%
	600 ms	50,65%	48,22%	43,36%
	1000 ms	47,10%	44,30%	37,57%
MFCC_E_Z media y desviación típica	60 ms	69,16%	69,72%	70,09%
	100 ms	69,91%	71,40%	67,10%
	300 ms	63,74%	58,69%	52,15%
	600 ms	55,33%	49,72%	44,11%
	1000 ms	52,15%	48,79%	41,87%
MFCC_E_Z media, desviación típica y skewness	60 ms	67,48%	67,10%	67,48%
	100 ms	66,92%	67,10%	62,43%
	300 ms	64,11%	57,94%	51,78%
	600 ms	54,21%	50,47%	45,23%
	1000 ms	54,95%	46,54%	40,93%
MFCC_D_E_Z media y desviación típica	60 ms	72,15%	71,59%	68,97%
	100 ms	69,72%	68,97%	65,23%
	300 ms	64,86%	59,63%	49,91%
	600 ms	56,82%	47,10%	44,30%
	1000 ms	54,02%	46,92%	39,81%

Tabla 6. Resultados de los experimentos por segmento basados en estadísticos a partir de los parámetros MFCC para la base de datos EMODB.

Como se puede observar en la tabla anterior, a medida que el tamaño del segmento generado aumenta, la precisión del reconocedor empeora. Esto puede ser debido a que al aumentar la longitud del segmento, disminuye el número de ejemplos de entrenamiento y dado que la base de datos es por sí de tamaño reducido, los modelos obtenidos no están correctamente entrenados, produciendo, por tanto, una disminución en la tasa de reconocimiento del sistema final.

Los mejores resultados se obtienen para segmentos de 60 o 100 milisegundos (compuestos por respectivamente, 6 o 10 tramas), lo que demuestra que al disminuir el número de muestras a tener en cuenta en nuestro experimento, tanto para el entrenamiento como para el posterior reconocimiento, el reconocedor resultante pierde precisión obteniéndose de esta forma peores resultados que los obtenidos en los experimentos de referencia. Del mismo modo, a medida que aumentamos el número de características en nuestros experimentos añadiendo a la media y la desviación típica, la curtosis y/o skewness, podemos observar cómo los resultados empeoran. El motivo de este comportamiento radica en que al mismo tiempo que disminuimos el número de muestras utilizadas en el entrenamiento aumentamos el número de parámetros a entrenar puesto que añadimos nuevas características a cada una de estas muestras. De esta forma, los modelos entrenados no generalizan bien. Debido a esto no es posible realizar el reconocimiento posterior con la misma precisión que en los experimentos de referencia.

Al igual que ocurría en los experimentos anteriores, los resultados con un mayor porcentaje de aciertos se corresponden con aquellos que incluyen las deltas (D) y la energía (E), obteniendo los mejores resultados para los experimentos MFCC_D_E_Z en los que se han calculado la media y desviación típica para segmentos de 60 milisegundos con una mezcla de 32 gaussianas (72,15% de aciertos) y una mezcla de 64 gaussianas (71,54% de aciertos). A partir de estos resultados, podemos observar también que a pesar de estar por debajo del 73,27% de precisión obtenido como referencia para un MFCC_D_E_Z con 23 filtros y una mezcla de 64 gaussianas, un porcentaje de aciertos de un 71,54% es un valor muy elevado teniendo en cuenta que contamos con un mayor número de parámetros que entrenar y un menor número de muestras para hacerlo.

A pesar de todo lo anteriormente señalado, es posible observar cómo en los casos en los que el número de parámetros de entrada al reconocedor es menor, exactamente en la utilización de MFCC_Z y MFCC_E_Z en los casos en los que únicamente se utiliza la media y la desviación típica para llevar a cabo los experimentos, los resultados son mejores que los obtenidos para un MFCC_Z y un MFCC_E_Z, respectivamente, en el caso de referencia. Siendo los mejores valores 71,40% para un MFCC_E_Z_media_y_desviación típica en segmentos de 100 milisegundos y 64 mezclas de gaussianas frente a un 60,19% para un MFCC_Z de 128 mezclas de gaussianas en el caso de referencia y un 67,10% MFCC_Z_media_y_desviación_típica en segmentos de 60 milisegundos y 64 mezclas de gaussianas frente a un 65,61% para un MFCC_E_Z de 128 mezclas de gaussianas en el caso de referencia.

Esto último, nos permite suponer que con un mayor número de muestras para nuestro entrenamiento sería posible que los resultados obtenidos al aplicar estos experimentos basados en estadísticos pudieran ser mejores que los obtenidos en el caso de referencia.

Una vez obtenido este conjunto de resultados, seleccionamos aquellos casos que nos proporcionaban mejores porcentajes de acierto y procedimos a repetir este mismo procedimiento añadiendo al conjunto de características el cálculo de medias y desviaciones típicas de la frecuencia fundamental del conjunto de tramas que forman un segmento. De este mismo modo se realizó otro experimento en el que se calculó a su vez la media y desviación típica no sólo de la frecuencia fundamental de estas tramas sino también de su máximo y mínimo en las mismas.

Los resultados de estos experimentos así como los obtenidos en las mismas condiciones pero en ausencia del cálculo del pitch, se muestran en la siguiente tabla.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas
MFCC_D_E_Z media y desviación típica	60 ms	72,15%	71,59%	68,97%	
MFCC_D_E_Z_pitch media y desviación típica	60 ms	67,48%	69,16%	68,60%	65,05%
MFCC_D_E_Z_pitch_pitchMax_pitchMin media y desviación típica	60 ms	64,11%	68,04%	66,36%	62,99%
MFCC_E_Z media y desviación típica	60 ms	69,16%	69,72%	70,09%	
MFCC_E_Z_pitch media y desviación típica	60 ms	64,30%	67,66%	68,60%	64,86%
MFCC_E_Z_pitch_pitchMax_pitchMin media y desviación típica	60 ms	62,24%	62,99%	63,74%	60,93%
MFCC_E_Z media y desviación típica	100 ms	69,91%	71,40%	67,10%	
MFCC_E_Z_pitch media y desviación típica	100 ms	64,11%	67,29%	65,61%	58,50%
MFCC_E_Z_pitch_pitchMax_pitchMin media y desviación típica	100 ms	62,24%	64,49%	63,74%	58,50%

Tabla 7. Resultados de los experimentos por segmento basados en estadísticos en presencia y ausencia de la frecuencia fundamental a partir de los parámetros MFCC para la base de datos EMODB.

Teniendo en cuenta los datos mostrados en la tabla anterior, podemos afirmar que al incluir el cálculo de la frecuencia fundamental en nuestros experimentos, así como su valor máximo y mínimo dentro de cada segmento, los resultados obtenidos empeoran. Esto puede ser debido a errores en la estimación del pitch y al incremento del número de parámetros a entrenar.

Experimentos basados en banco de filtros

El método utilizado para realizar el siguiente conjunto de experimentos sigue el proceso explicado en el apartado llamado *Método basado en banco de filtros* del tercer capítulo. En éste, tras agrupar los parámetros acústicos a nivel de trama en segmentos de un tamaño determinado, se procede a calcular sus periodogramas, sobre los que, finalmente, se aplica un banco de filtros para reducir su dimensionalidad y compactar la información contenida en los mismos.

En la siguiente tabla se muestran los resultados de este proceso.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
MFCC_Z	60 ms	38,69%	42,99%	42,99%	43,36%
	100 ms	47,10%	48,79%	51,78%	47,48%
	300 ms	49,16%	51,59%	48,41%	39,81%
	600 ms	51,78%	48,22%	42,43%	34,77%
	1000 ms	52,90%	47,48%	42,99%	34,39%
MFCC_E_Z	60 ms	46,54%	50,47%	53,46%	51,78%
	100 ms	53,08%	55,70%	56,26%	51,96%
	300 ms	53,64%	55,51%	51,59%	46,36%
	600 ms	54,39%	50,84%	45,98%	40,56%
	1000 ms	56,45%	50,65%	44,86%	37,20%
MFCC_D_Z	60 ms	48,79%	48,22%	48,22%	49,91%
	100 ms	49,16%	53,08%	49,35%	46,36%
	300 ms	52,15%	48,97%	46,73%	39,81%
	600 ms	51,96%	48,60%	42,99%	39,07%
	1000 ms	49,35%	46,17%	40,37%	35,14%
MFCC_D_E_Z	60 ms	53,46%	53,08%	54,02%	53,83%
	100 ms	55,33%	56,64%	54,21%	50,09%
	300 ms	54,02%	52,15%	48,22%	42,06%
	600 ms	54,21%	51,96%	45,23%	37,38%
	1000 ms	50,84%	45,23%	40,19%	35,70%

Tabla 8. Resultados de los experimentos por segmento basados en bancos de filtros a partir de los parámetros MFCC para la base de datos EMODB.

Una vez observados los resultados presentados en la tabla anterior, se puede concluir que los experimentos basados en banco de filtros producen peores resultados que los basados en estadísticos. Por ejemplo, el mejor valor para el caso de un experimento basado en banco de filtros, 56,64%, se consigue para un MFCC_D_E_Z de 100 milisegundos con una mezcla de 64 gaussianas, mientras que en el mismo caso para un experimento basado en estadísticos en los que se ha calculado únicamente la media y la desviación típica es de 68,97%.

Al igual que ocurría en el apartado anterior, es posible observar cómo los resultados empeoran a medida que aumentamos el tamaño de los segmentos. Sin embargo, en este caso los mejores resultados corresponden a segmentos de 100 y 300 milisegundos en lugar de segmentos de 60 y 100 milisegundos.

Experimentos basados en estadísticos y banco de filtros

En este último apartado combinamos las dos técnicas explicadas en los apartados anteriores. A la vista de los resultados obtenidos con anterioridad, únicamente se han tenido en cuenta la media y la desviación típica de los diferentes segmentos como parámetros para la parte de los experimentos basados en estadísticos, mientras que la parte de los experimentos basados en banco de filtros se mantiene íntegra.

Los resultados de estos experimentos se muestran a continuación.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
MFCC_Z	60 ms	50,47%	51,78%	54,58%	51,78%
	100 ms	53,08%	55,33%	54,58%	50,28%
	300 ms	55,70%	53,27%	47,10%	40,00%
	600 ms	51,59%	47,48%	43,93%	37,20%
	1000 ms	51,21%	47,29%	40,75%	36,26%
MFCC_E_Z	60 ms	55,89%	60,75%	63,18%	58,32%
	100 ms	57,38%	60,00%	60,93%	58,69%
	300 ms	59,63%	58,50%	52,34%	45,05%
	600 ms	54,21%	51,21%	44,11%	38,88%
	1000 ms	55,89%	47,85%	41,31%	33,64%
MFCC_D_Z	60 ms	53,08%	51,96%	52,15%	50,09%
	100 ms	56,26%	54,21%	52,71%	48,41%
	300 ms	57,01%	54,21%	48,41%	41,50%
	600 ms	51,59%	47,85%	42,06%	36,64%
	1000 ms	48,60%	44,11%	37,94%	35,51%
MFCC_D_E_Z	60 ms	59,25%	60,75%	60,37%	57,76%
	100 ms	56,64%	58,69%	58,32%	51,96%
	300 ms	60,00%	57,94%	49,72%	43,36%
	600 ms	56,26%	49,35%	42,62%	37,01%
	1000 ms	49,16%	45,98%	39,07%	32,15%

Tabla 9. Resultados de los experimentos por segmento basados en estadísticos y bancos de filtros a partir de los parámetros MFCC para la base de datos EMODB.

Como se puede apreciar al observar la tabla anterior, este conjunto de experimentos nos genera mejores resultados que aquellos obtenidos en los basados en banco de filtros. Sin embargo, los experimentos basados en estadísticos siguen proporcionándonos los mejores resultados dentro del conjunto de experimentos por segmento. Como ejemplo de esto último mostraremos el mismo caso que utilizamos en el apartado anterior para comparar los experimentos basados en bancos de filtros con los basados en estadísticos. Para un experimento MFCC_D_E_Z de 100 milisegundos con una mezcla de 64 gaussianas obtuvimos un resultado de 56,64% para el caso del experimento basado en banco de filtros, 68,97% para el del experimento basado en estadísticos en el que se ha incluido únicamente la media y la desviación típica de los segmentos y de 58,69% para el experimento en el que se han mezclado las dos técnicas anteriores.

Con respecto a los datos obtenidos a partir de los experimentos en los que se han calculado los parámetros MFCC, podemos concluir que los experimentos por trama nos aportan mejores resultados que aquellos basados en segmentos. Esto puede deberse, como se ha comentado con anterioridad, a que al combinar los datos de los diferentes segmentos obtenemos un conjunto de ejemplos de entrenamiento de menor tamaño, por lo que el entrenamiento de los diferentes modelos de gaussianas puede ser peor.

Experimentos con los parámetros FFLFBE

En esta sección se van a llevar a cabo los mismos experimentos que en la inmediatamente anterior, siendo la diferencia entre ambas que en este caso se utilizan como base los parámetros FFLFBE, explicados en el apartado *Frequency-Filtered Logarithmic Filter Bank Energies (FFLFBE)* del tercer capítulo en lugar de los MFCC.

Al igual que ocurría en los anteriores experimentos, vamos a distinguir dos tipos: por trama y por segmento.

Por trama

En el caso de los experimentos por trama para los parámetros FFLFBE, realizamos cuatro tipos de experimentos, dos en los que se utilizaban 23 bandas en el banco de filtros utilizado por HTK y otros dos en los que se utilizaban 25. El motivo de estos experimentos está en el modo de calcular los parámetros FFLFBE. Tal y como se ha explicado en el tercer capítulo de este proyecto, los coeficientes FFLFBE $\left(X_t^{lf}\right)$ de una trama t cualquiera, se calcularían de la siguiente forma:

$$X_t^{lf}(0) = X_t^l(1)$$

$$X_t^{lf}(m) = X_t^l(m+1) - X_t^l(m-1), \qquad 0 < m \leq N_b - 2$$

$$X_t^{lf}(N_b - 1) = -X_t^l(N_b - 2)$$

Al observar las ecuaciones anteriores en las que X_t^l son las log-energías en banda y N_b es el número de bandas totales, se puede apreciar que tanto el primer elemento de las log-energías en banda ($X_t^l(0)$) como el último ($X_t^l(N_b - 1)$) no se tienen en cuenta al calcular los coeficientes de estos parámetros. Debido a esto pensamos que añadiendo dos filtros más al banco de filtros, utilizando un valor de 25 en lugar de 23 en el parámetro NUMCHAN de la configuración de HTK sería posible mejorar los resultados correspondientes a estos experimentos, puesto que conservaríamos una mayor cantidad de información útil en los mismos obteniendo en el primer caso 23 coeficientes y 21 en el segundo caso.

El segundo tipo de experimentos que se realizaron con estas características consistía en restar, o no, la media total de cada coeficiente a los coeficientes calculados en cada una de las tramas. La nomenclatura Z se corresponde con coeficientes en los que la media es 0 por lo que para ser completamente rigurosos era necesario sustraer la media a estos coeficientes. Sin embargo, antes de establecer este proceso era necesario comprobar la validez del mismo por lo que, para ello, llevamos a cabo dos tipos de experimentos con 23 y 25 canales. En uno de ellos se eliminaba la media de estos coeficientes, mientras que en el otro no se modificaba.

A continuación se muestra una tabla con los datos obtenidos en los cuatro casos anteriores.

			Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
			Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_Z	23 filtros	Sin eliminar media	58,69%	60,19%
		Eliminando media	60,75%	60,19%
	25 filtros	Sin eliminar media	57,20%	55,14%
		Eliminando media	57,01%	57,38%
FFLFBE_E_Z	23 filtros	Sin eliminar media	61,87%	63,36%
		Eliminando media	63,36%	63,74%
	25 filtros	Sin eliminar media	57,38%	59,44%
		Eliminando media	57,38%	58,88%
FFLFBE_D_Z	23 filtros	Sin eliminar media	65,79%	67,85%
		Eliminando media	66,17%	66,17%
	25 filtros	Sin eliminar media	63,55%	64,49%
		Eliminando media	63,93%	63,36%
FFLFBE D_E_Z	23 filtros	Sin eliminar media	67,66%	68,60%
		Eliminando media	68,22%	68,97%
	25 filtros	Sin eliminar media	62,80%	64,11%
		Eliminando media	66,17%	68,04%
FFLFBE D_A_Z	23 filtros	Sin eliminar media	68,41%	69,91%
		Eliminando media	70,84%	72,52%
	25 filtros	Sin eliminar media	64,67%	64,67%
		Eliminando media	65,98%	65,98%
FFLFBE_A_D_E_Z	23 filtros	Sin eliminar media	72,34%	71,96%
		Eliminando media	73,08%	72,90%
	25 filtros	Sin eliminar media	69,91%	68,79%
		Eliminando media	69,53%	70,47%

Tabla 10. Resultados de los experimentos por trama con 23 y 25 filtros a partir de los parámetros FFLFBE para la base de datos EMODB.

Al tratarse de un conjunto de experimentos pensado para establecer el número de filtros y la presencia o ausencia de la media en estos parámetros, únicamente se tuvieron en cuenta dos tipos de mezclas de gaussianas, una de 64 gaussianas y otra de 128. Otra de las razones para utilizar estas dos mezclas de gaussianas, era poder comparar estos resultados con los datos obtenidos en los experimentos de referencia llevados a cabo a partir de los parámetros MFCC.

Una vez realizado un estudio de los datos aquí mostrados, se puede concluir que con el número de filtros igual a 23 los experimentos producen mejores resultados en todos los casos. Además, comparando los casos en los que hemos eliminado la media con aquellos en los que no se ha modificado ningún aspecto de los mismos, se puede observar cómo en la mayoría de estos experimentos los mejores resultados se obtienen a partir de las tramas cuya media es 0. De esta forma, procedimos a fijar el número de filtros a 23 y a eliminar la media de cada uno de los parámetros de las diversas tramas.

En la siguiente tabla se muestran los valores de los experimentos que nos servirán como referencia para los parámetros FFLFBE.

	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas
FFLFBE_Z	60,75%	60,19%
FFLFBE_E_Z	63,36%	63,74%
FFLFBE_D_Z	66,17%	66,17%
FFLFBE_E_D _Z	68,22%	68,97%
FFLFBE_D_A _Z	70,84%	72,52%
FFLFBE_E_D_A _Z	73,08%	72,90%

Tabla 11. Resultados de los experimentos por trama utilizando 23 filtros a partir de los parámetros FFLFBE para la base de datos EMODB.

Al igual que ocurría en el caso de los parámetros MFCC, la presencia de la energía así como la de los deltas aumenta notablemente el número de muestras correctamente reconocidas. La combinación de ambas características produce, a su vez, unos resultados mejores.

Con respecto a las aceleraciones, podemos constatar que, en este caso, la presencia de las mismas produce resultados más determinantes que los obtenidos a partir de estos mismos experimentos realizados sobre los parámetros MFCC.

Como puede observarse comparando la tabla de referencia de los parámetros MFCC con la con la tabla anterior, los coeficientes FFLFBE consiguen resultados ligeramente por debajo de los MFCC.

Por segmento

Siguiendo el mismo procedimiento que en el caso de los parámetros MFCC, en este apartado dividimos las tramas que contienen los coeficientes FFLFBE en segmentos de 100, 300, 600 o 1000 milisegundos y realizamos una serie de modificaciones a los mismos antes de llevar a cabo el entrenamiento del reconocedor de emociones y el propio reconocimiento de las mismas.

A diferencia del caso anterior y como consecuencia de las conclusiones obtenidas a partir de la observación de los resultados de los experimentos sobre los parámetros MFCC, únicamente hemos llevado a cabo un conjunto seleccionado de experimentos basados en estadísticos sobre los coeficientes FFLFBE.

Experimentos basados en estadísticos

En este conjunto de experimentos, en los distintos segmentos obtenidos se calculan la media y desviación típica de los parámetros contenidos en dichos segmentos. A la vista de los resultados obtenidos para el caso de los parámetros MFCC, los experimentos que incluían la skewnes y/o curtosis no se han llevado a cabo para estos coeficientes.

A continuación se muestran los resultados obtenidos.

		Modelos de 32 mezclas de gaussianas	**Modelos de 64 mezclas de gaussianas**	**Modelos de 128 mezclas de gaussianas**
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_Z media y desviación típica	100 ms	66,17%	68,41%	64,30%
	300 ms	62,06%	60,75%	51,78%
	600 ms	57,94%	49,16%	43,55%
	1000 ms	53,46%	45,23%	38,69%
FFLFBE_E_Z media, desviación típica	100 ms	70,47%	67,10%	63,55%
	300 ms	65,79%	65,23%	56,82%
	600 ms	58,32%	52,34%	44,30%
	1000 ms	54,02%	48,22%	44,30%
FFLFBE_E_D_Z media, desviación típica	100 ms	71,40%	70,28%	66,92%
	300 ms	67,48%	64,86%	55,33%
	600 ms	59,63%	54,21%	46,36%
	1000 ms	54,02%	47,10%	40,00%

Tabla 12. Resultados de los experimentos por segmento basados en estadísticos a partir de los parámetros FFLFBE para la base de datos EMODB.

En la tabla anterior puede observarse cómo a medida que aumenta el tamaño de los segmentos utilizados, los resultados obtenidos empeoran progresivamente. El motivo de esto, al igual que en el caso anterior, se basa en que al aumentar el tamaño del segmento se disminuyen el número de ejemplos utilizados en el entrenamiento de los modelos de emociones, por lo que la precisión del sistema global es menor.

Otro hecho que puede percibirse es que las mejores tasas de reconocimiento corresponden a una mezcla de 32 gaussianas en la gran mayoría de los casos. Además, teniendo en cuenta lo observado en los experimentos por trama, podemos confirmar que también en este caso la presencia tanto de la energía como de las deltas mejoran considerablemente los resultados obtenidos.

En la siguiente tabla se muestra la última tanda de experimentos que se llevó a cabo para estos parámetros en la base de datos EMO-DB. En estos, hemos añadido el pitch de la señal de audio para comprobar su efecto en este caso.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_E_Z media y desviación típica	100 ms	69,35%	66,73%	61,12%	
FFLFBE E_Z_pitch media y desviación típica	100 ms	70,65%	67,66%	64,86%	59,44%
FFLFBE_E_Z_pitch_pitchMax_pitchMin media y desviación típica	100 ms	67,85%	66,54%	64,86%	56,82%
FFLFBE_D_E_Z media y desviación típica	100 ms	72,52%	69,91%	66,73%	
FFLFBE_D_E_Z_pitch media y desviación típica	100 ms	70,65%	67,66%	64,86%	59,44%
FFLFBE_E_Z_pitch_pitchMax_pitchMin media y desviación típica	100 ms	67,85%	66,54%	64,86%	56,82%
FFLFBE_D_E_Z media y desviación típica	300 ms	69,53%	62,43%	54,95%	
FFLFBE_D_E_Z_pitch media y desviación típica	300 ms	64,30%	62,43%	54,39%	48,41%
FFLFBE_D_E_Z_pitch_pitchMax_pitchMin media y desviación típica	100 ms	61,87%	59,25%	53,46%	46,73%

Tabla 13. Resultados de los experimentos por segmento basados en estadísticos en presencia y ausencia de la frecuencia fundamental a partir de los parámetros FFLFBE para la base de datos EMODB.

Observando los datos anteriores podemos concluir que, al igual que ocurría en el caso de los coeficientes MFCC, añadir el pitch a los vectores de parámetros supone un ligero empeoramiento de los resultados obtenidos en el reconocimiento de emociones.

Resumen de los resultados de la base de datos EMODB

A continuación se muestra una gráfica que contiene un resumen de los mejores resultados obtenidos a partir de los experimentos realizados con el conjunto de muestras acústicas contenidas en la base de datos EMODB.

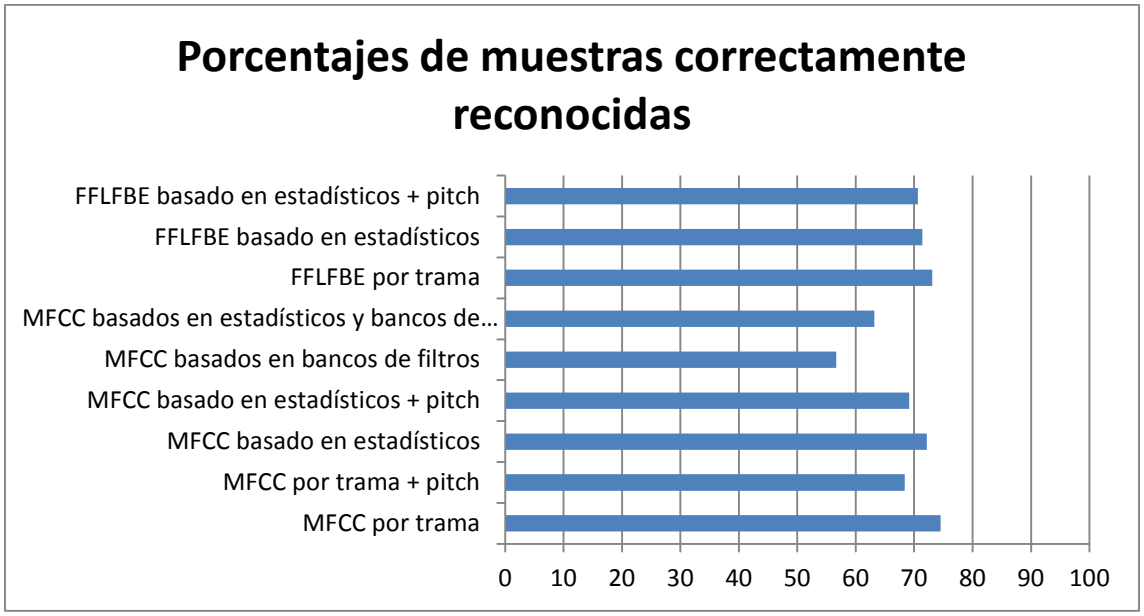

Figura 14. Resumen de los resultados en porcentaje obtenidos para la base de datos EMODB.

Como puede observarse en la figura anterior, los mejores resultado obtenidos para el conjunto de experimentos realizados en la base de datos EMODB se corresponden con los experimentos MFCC por trama, seguidos de los experimentos FFLFBE por trama y los MFCC basados en estadísticos.

Resultados de la base de datos SAVEE

En este apartado se van a mostrar los resultados obtenidos a partir de los experimentos llevados a cabo sobre la base de datos SAVEE. Al igual que ocurría en la sección anterior, se van a efectuar dos tipos de experimentos: experimentos basados en los parámetros MFCC y experimentos basados en los parámetros FFLFBE.

Experimentos con los parámetros MFCC

Tal como se ha explicado para la base de datos EMODB en este mismo capítulo, en esta sección se van a mostrar los resultados obtenidos a partir de los experimentos llevados a cabo sobre los parámetros MFCC.

Se han desarrollado dos tipos de experimentos: por trama y por segmento.

Por trama

Al igual que con la base de datos EMODB, en este subapartado se van a presentar los resultados obtenidos a partir de los experimentos de referencia o baseline. Tras llevarse a cabo la división de la señal acústica en un conjunto de tramas, se realizaron una serie de experimentos que incluían distintas combinaciones de parámetros, con el fin de descubrir cómo estas combinaciones afectan al reconocimiento de emociones efectuado sobre esta nueva base de datos.

Tanto la nomenclatura de los experimentos y resultados, como las funciones utilizadas en este proceso son las mismas que en el caso de la base de datos EMODB.

Los datos recogidos en esta primera tanda de experimentos son los siguientes:

	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
	Porcentaje de muestras correctamente reconocidas	Porcentaje de muestras correctamente reconocidas
MFCC_Z	62,50%	66,04%
MFCC_E_Z	66,88%	67,92%
MFCC_D_Z	60,62%	61,88%
MFCC_E_D _Z	65,62%	65,00%
MFCC_E_D_A _Z	64,52%	65,00%

Tabla 14. Resultados de los experimentos por trama a partir de los parámetros MFCC para la base de datos SAVEE.

En la tabla anterior puede observarse cómo los modelos con una mezcla de 128 gaussianas ofrecen mejores resultados que los modelos con una mezcla de 64 gaussianas. Este mismo patrón se repite también para el mismo conjunto de experimentos en el caso de la base de datos EMODB, sin embargo, los porcentajes obtenidos a partir del reconocimiento de emociones de la misma son superiores a los conseguidos en la base de datos SAVEE llegando a obtener un 74,50% para el caso MFCC_E_D_A_Z de la base de datos EMODB frente a un 65,00% de la base de datos SAVEE.

Con respecto a los valores recopilados en esta tanda de experimentos podemos señalar que al igual que ocurría en el caso anterior, la energía (E) es un parámetro fundamental para este proceso. La presencia de este parámetro en dichos experimentos supone una mejora de los resultados obtenidos llegando a alcanzar un 65,62% de muestras correctamente reconocidas para el caso MFCC_E_D_Z frente a un 60,62% para el caso MFCC_D_Z.

A diferencia de lo ocurrido en los experimentos comentados anteriormente, tanto la presencia de las deltas (D) como la de las aceleraciones (A) resulta menos crítica, es decir, no se obtienen resultados mucho mejores incluyendo estos parámetros en el proceso.

Por segmento

Siguiendo el mismo proceso que en experimentos anteriores, las tramas obtenidas a partir de las muestras de audio se dividen en segmentos de 100, 300, 600 o 1000 milisegundos y, tras realizar una serie de modificaciones a los parámetros MFCC previamente calculados, se procede con el entrenamiento del reconocedor de emociones y el propio reconocimiento de las mismas.

De igual forma, llevamos a cabo dos tipos de experimentos: basados en estadísticos y basados en bancos de filtros.

Experimentos basados en estadísticos

En este conjunto de experimentos se lleva a cabo el cálculo de la media y desviación típica para los segmentos previamente mencionados. Los datos proporcionados por estos cálculos se utilizan como características de entrada tanto para el entrenamiento como para el reconocimiento de las emociones. Atendiendo a los resultados obtenidos para la base de datos EMODB se han eliminado los experimentos en los que se incluía el cálculo de la skewnes y curtosis.

A continuación se muestra una tabla que recoge los resultados de estos experimentos.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
		Porcentaje muestras reconocidas	Porcentaje muestras reconocidas	Porcentaje muestras reconocidas
MFCC_Z media y desviación típica	60 ms	56,25%	61,25%	61,04%
	100 ms	58,54%	60,00%	61,25%
	300 ms	53,54%	55,00%	49,17%
	600 ms	42,92%	47,71%	39,38%
	1000 ms	40,80%	36,46%	31,04%
MFCC_E_Z media y desviación típica	60 ms	60,83%	61,25%	62,08%
	100 ms	59,58%	61,25%	62,50%
	300 ms	56,25%	54,38%	49,38%
	600 ms	46,04%	45,83%	40,21%
	1000 ms	43,12%	39,17%	33,96%
MFCC_E_D_Z media y desviación típica	60 ms	59,79%	58,75%	57,08%
	100 ms	56,46%	59,38%	57,71%
	300 ms	49,79%	51,46%	45,21%
	600 ms	43,75%	40,83%	34,17%
	1000 ms	42,50%	38,96%	30,00%
MFCC_D_Z media y desviación típica	60 ms	54,17%	56,25%	55,83%
	100 ms	53,75%	55,62%	54,17%
	300 ms	48,12%	43,96%	39,79%
	600 ms	38,96%	42,71%	34,79%
	1000 ms	42,08%	36,04%	30,21%

Tabla 15. Resultados de los experimentos por segmento basados en estadísticos a partir de los parámetros MFCC para la base de datos SAVEE.

Atendiendo a los datos recogidos en la tabla anterior puede confirmarse que, al igual que ocurría en la base de datos EMODB, la energía de las muestras es un parámetro fundamental para el reconocimiento de emociones obteniéndose los mejores porcentajes para los conjuntos de características en los que ésta ha sido incluida. Sin embargo, a diferencia de lo que ocurría en la base de datos anteriormente mencionada, la presencia de las deltas (D) no mejora los porcentajes obtenidos únicamente con el parámetro de energía.

Es necesario también señalar que a medida que el tamaño del segmento aumenta el porcentaje de muestras correctamente reconocidas disminuye. Al igual que en el caso anterior esto es debido a que un mayor tamaño del segmento supone un menor número de muestras para el entrenamiento del reconocedor, por lo que la precisión del mismo tiende a disminuir. Como muestra de lo anterior podemos comentar que los mejores resultados se obtienen para segmentos de 60 o 100 milisegundos.

A la vista de los resultados obtenidos en la base de datos anterior no se han incluido en este caso los experimentos que contenían el cálculo de la curtosis y/o skewness debido a que dichos resultados proporcionaban peores porcentajes que los realizados únicamente teniendo en cuenta la media y la desviación típica.

A diferencia de lo que ocurría en la base de datos EMODB, los mejores resultados de estos experimentos basados en estadísticos se corresponden con aquellos que incluyen únicamente la energía (E), obteniendo el mayor porcentaje para el experimento MFCC_E_Z en el que se ha calculado la media y desviación típica para segmentos de 100 milisegundos con una mezcla de 128 gaussianas (62,50%) y el MFCC_E_Z para segmentos de 60 milisegundos con una mezcla de 128 gaussianas en el que se ha calculado la media y la desviación típica (62,08%).

En este caso, los resultados obtenidos en los experimentos basados en segmentos no superan el porcentaje proporcionado por los experimentos de referencia, por lo que podemos concluir que un mayor número de características así como un menor número de muestras dan lugar a peores resultados.

Una vez concluido el conjunto de experimentos anterior, seleccionamos aquellos casos que nos proporcionaban mejores porcentajes de acierto y llevamos a cabo este mismo procedimiento añadiendo a estos la frecuencia fundamental del conjunto de tramas que forman un segmento. Con el objetivo de ser coherentes con respecto a la base de datos EMODB, se realizó otro experimento en el que se calculó la media no sólo de la frecuencia fundamental de estas tramas sino también de su máximo y mínimo además de la media y desviación típica de los parámetros MFCC de estas tramas.

Los resultados de estos experimentos así como los obtenidos en las mismas condiciones pero teniendo en cuenta la frecuencia fundamental, se muestran en la siguiente tabla.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje muestras reconocidas	Porcentaje muestras reconocidas	Porcentaje muestras reconocidas	Porcentaje muestras reconocidas
MFCC_Z media y desviación típica	60 ms	56,25%	61,25%	61,04%	
MFCC_Z_pitch media y desviación típica	60 ms	57,71%	61,88%	62,08%	60,83%
MFCC_Z_pitch_pitch Max_pitchMin media y desviación típica	60 ms	58,54%	62,71%	61,46%	59,17%
MFCC_Z media y desviación típica	100 ms	58,54%	60,00%	61,25%	
MFCC_Z_pitch media y desviación típica	100 ms	57,71%	59,38%	60,21%	58,12%
MFCC_Z_pitch_pitch Max_pitchMin media y desviación típica	100 ms	56,88%	59,58%	61,67%	58,12%
MFCC_E_Z media y desviación típica	60 ms	60,83%	61,25%	62,08%	
MFCC_E_Z_pitch media y desviación típica	60 ms	61,88%	63,96%	64,17%	62,08%
MFCC_E_Z_pitch_pit chMax_pitchMin media y desviación típica	60 ms	64,58%	63,12%	62,29%	59,58%
MFCC_E_Z media y desviación típica	100 ms	59,58%	61,25%	62,50%	
MFCC_E_Z_pitch media y desviación típica	100 ms	60,21%	60,42%	64,38%	59,17%
MFCC_E_Z_pitch_pit chMax_pitchMin media y desviación típica	100 ms	58,54%	63,54%	65,21%	62,08%

Tabla 16. Resultados de los experimentos por segmento basados en estadísticos en presencia y ausencia de la frecuencia fundamental a partir de los parámetros MFCC para la base de datos SAVEE.

A partir de los datos anteriores, podemos observar que al incluir el cálculo de la frecuencia fundamental, así como los valores máximo y mínimo de la misma, los resultados obtenidos mejoran, obteniendo los mejores resultados para las mezclas de 128 gaussianas y los experimentos que incluyen la energía como característica (MFCC_E_Z). Por tanto, el mejor resultado obtenido es un 65,21% de aciertos que se obtienen para un MFCC_E_Z en el que se

ha añadido tanto la frecuencia fundamental como su máximo y mínimo para una mezcla de 128 gaussianas y segmentos de 100 milisegundos.

Puesto que los resultados obtenidos son mejores que los proporcionados por los experimentos basados en estadísticos en los que el pitch no se ha incluido, podemos afirmar, de esta forma, que la frecuencia fundamental es un parámetro importante para el reconocimiento de emociones en la base de datos SAVEE.

Experimentos basados en banco de filtros

El siguiente apartado utiliza el método basado en banco de filtros explicado en el capítulo número tres. Este método se basa en el cálculo del periodograma de los diferentes segmentos que han sido obtenidos tras agrupar los parámetros acústicos a nivel de trama, sobre los que se aplicará un banco de filtros con el fin de reducir su dimensionalidad y compactar la información contenida en los mismos.

En la siguiente tabla se muestran los resultados del proceso anteriormente explicado.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
MFCC_Z	60 ms	42,08%	44,38%	49,17%	47,29%
	100 ms	41,46%	42,71%	45,62%	45,62%
	300 ms	42,50%	39,58%	38,33%	32,08%
	600 ms	35,21%	35,21%	30,62%	28,12%
	1000 ms	37,08%	33,75%	30,62%	30,42%
MFCC_E_Z	60 ms	40,21%	46,88%	50,83%	52,92%
	100 ms	46,67%	49,38%	49,58%	47,29%
	300 ms	43,12%	41,04%	41,25%	33,75%
	600 ms	40,00%	36,88%	33,96%	26,67%
	1000 ms	37,08%	31,67%	31,88%	26,67%
MFCC_D_Z	60 ms	41,04%	45,83%	48,96%	47,29%
	100 ms	42,08%	44,38%	45,21%	41,04%
	300 ms	36,67%	39,38%	35,00%	28,54%
	600 ms	38,33%	32,71%	32,92%	25,62%
	1000 ms	35,21%	33,33%	28,75%	24,58%
MFCC_D_E_Z	60 ms	42,71%	45,21%	47,50%	45,00%
	100 ms	40,00%	45,42%	45,42%	43,33%
	300 ms	38,75%	42,08%	38,75%	29,58%
	600 ms	35,83%	37,08%	30,83%	26,04%
	1000 ms	32,71%	32,71%	28,12%	25,21%

Tabla 17. Resultados de los experimentos por segmento basados en bancos de filtros a partir de los parámetros MFCC para la base de datos SAVEE.

Comparando los resultados conseguidos a partir de las muestras acústicas procedentes de la base de datos SAVEE con aquellas obtenidas a partir de la base de datos EMODB, podemos observar que aunque no hay grandes diferencias entre ellos, los valores correspondientes a esta segunda base de datos son ligeramente mejores llegando a obtener como máximo un 56,26% en un experimento MFCC_E_Z con una mezcla de 128 gaussianas y segmentos de 100 milisegundos frente a un 52,92% para un experimento MFCC_E_Z con una mezcla de 256 gaussianas y segmentos de 60 milisegundos.

Al igual que ocurría en los experimentos realizados con la base de datos anterior, los experimentos basados en banco de filtros producen peores resultados que los basados en estadísticos. Por ejemplo, el mejor valor para el caso de un experimento basado en banco de filtros para un MFCC_E_Z de 60 milisegundos con una mezcla de 128 gaussianas se obtiene un 50,83% de muestras correctamente reconocidas, mientras que el mismo experimento pero basado en estadísticos produce un 62,08% de muestras correctamente reconocidas.

Una constante en todos estos experimentos es el empeoramiento que sufren los resultados a medida que aumentamos el tamaño de los segmentos.

Teniendo en cuenta que los resultados de los experimentos basados en estadísticos y bancos de filtros de la base de datos EMODB eran peores que los obtenidos en los otros dos experimentos, tomamos la decisión de no repetirlos para esta base de datos.

Experimentos con los parámetros FFLFBE

En esta sección se van a llevar a cabo los mismos experimentos que en la sección anterior utilizando en este caso los parámetros FFLFBE en lugar de los MFCC.

Al igual que ocurría en los anteriores experimentos, vamos a distinguir dos tipos: por trama y por segmento.

Por trama

En el caso de los experimentos por trama, a partir de los datos obtenidos con los parámetros FFLFBE de la base de datos EMODB establecimos que el número de bandas que mejor resultados nos ofrecían era 23. El cálculo de los parámetros FFLFBE se realiza de igual forma que en la base de datos anterior. Este proceso se ha explicado en el tercer capítulo de este proyecto.

Sin embargo, el segundo tipo de experimentos que se realizaron en la base de datos EMODB sí fue repetido para la base de datos SAVEE. Las características de estos experimentos consistían en restar, o no, la media total de cada coeficiente a los coeficientes calculados en cada una de las tramas. La nomenclatura utilizada es la misma que en el resto de los casos.

A continuación se muestra una tabla con los datos obtenidos.

		Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_Z	Sin eliminar media	74,79%	76,67%
	Eliminando media	71,88%	71,88%
FFLFBE_E_Z	Sin eliminar media	75,21%	78,33%
	Eliminando media	68,96%	72,29%
FFLFBE_D_Z	Sin eliminar media	72,08%	73,75%
	Eliminando media	69,79%	72,71%
FFLFBE D_E_Z	Sin eliminar media	71,25%	73,12%
	Eliminando media	70,00%	71,04%
FFLFBE D_A_Z	Sin eliminar media	66,25%	71,67%
	Eliminando media	68,12%	70,42%
FFLFBE_A_D_E_Z	Sin eliminar media	70,62%	71,46%
	Eliminando media	67,50%	70,62%

Tabla 18. Resultados de los experimentos por trama eliminando o no la media a partir de los parámetros FFLFBE para la base de datos SAVEE.

El objetivo de la tabla de datos anterior era el de establecer si la presencia o ausencia de la media en estos parámetros afectaba a los resultados generados por los mismos.

Una comparación de los valores anteriores nos permite observar cómo, a diferencia de lo que ocurría con la base de datos EMODB, en la mayoría de los experimentos los mejores resultados se obtienen a partir de las tramas cuya media no ha sido eliminada. Debido a esto, se estableció que la media no sería eliminada para los siguientes experimentos.

En la siguiente tabla se muestran los valores de los experimentos que nos servirán como referencia para los parámetros FFLFBE.

	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_Z	74,79%	76,67%
FFLFBE_E_Z	75,21%	78,33%
FFLFBE_D_Z	72,08%	73,75%
FFLFBE D_E_Z	71,25%	73,12%
FFLFBE D_A_Z	66,25%	71,67%
FFLFBE_A_D_E_Z	70,62%	71,46%

Tabla 19. Resultados de los experimentos por trama a partir de los parámetros FFLFBE para la base de datos SAVEE.

Al igual que ocurría en el caso de los parámetros MFCC, la presencia de la energía aumenta notablemente el número de muestras correctamente reconocidas, sin embargo, la presencia del resto de parámetros analizados (deltas, aceleraciones y combinaciones de estos entre ellos o con la energía) supone un empeoramiento de los mismos.

Si realizamos una comparativa teniendo en cuenta los mejores resultados de los cuatro experimentos por trama, previamente efectuados, obtendríamos la siguiente tabla.

		Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
EMODB	MFCC_E_D_A _Z	74,21%	74,50%
EMODB	FFLFBE_E_D_A _Z	73,08%	72,90%
SAVEE	MFCC_E_Z	66,88%	67,92%
SAVEE	FFLFBE_E_Z	75,21%	78,33%

Tabla 20. Resumen de los mejores resultados de los experimentos por trama para las bases de datos EMODB y SAVEE.

Una comparación de los mejores valores de ambas bases de datos mostrará que existen grandes diferencias entre ellas. No sólo las características que producen los mejores resultados en cada una de ellas son distintas (energía, deltas y aceleraciones para EMODB y únicamente energía en el caso de SAVEE), sino que, además, dependiendo de la base de datos los resultados obtenidos por unos determinados parámetros mejores que los otros, es decir, en el caso de la base de datos EMODB son los parámetros MFCC los que mejores datos producen (74,50%), mientras que en el caso de la base de datos SAVEE son los parámetros FFLFBE (78,33%) los que proporcionan un porcentaje por encima de todos los anteriores.

Esta diferencia entre ambas bases de datos puede deberse a que en ellas se utilizan diferentes idiomas (alemán e inglés) con actores de dichas nacionalidades. Esto produce grandes diferencias en las muestras acústicas lo cual, tal como puede observarse en las tablas anteriores, genera grandes diferencias en los resultados obtenidos.

Por segmento

Tal como se ha explicado con anterioridad, para llevar a cabo los experimentos basados en segmentos las tramas que contienen los coeficientes FFLFBE se dividen en segmentos de 100, 300, 600 o 1000 milisegundos y, sobre ellos, se realizan una serie de cálculos o transformaciones antes de proceder al entrenamiento del reconocedor de emociones y al propio reconocimiento de las mismas.

Al igual que ocurría con la base de datos anterior, únicamente se ha llevado a cabo un conjunto seleccionado de experimentos basados en estadísticos sobre los coeficientes FFLFBE.

Experimentos basados en estadísticos

En los experimentos llevados a cabo sobre los parámetros FFLFBE y basados en estadísticos, se ha calculado la media y desviación típica sobre los distintos segmentos obtenidos por el procedimiento previamente mencionado. Teniendo en cuenta los resultados obtenidos para el caso de los parámetros MFCC, los experimentos que incluían la skewnes y/o curtosis no se han realizado para estos coeficientes.

En la siguiente tabla se muestran los resultados obtenidos a través de dicho proceso.

		Modelos de 32 mezclas de gaussianas	**Modelos de 64 mezclas de gaussianas**	**Modelos de 128 mezclas de gaussianas**
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_Z media y desviación típica	60 ms	63,96%	68,12%	72,08%
	100 ms	66,04%	69,38%	71,88%
	300 ms	66,25%	65,62%	65,42%
	600 ms	67,71%	63,33%	57,71%
	1000 ms	63,54%	58,54%	48,12%
FFLFBE_E_Z media, desviación típica	60 ms	64,17%	67,92%	72,92%
	100 ms	63,96%	66,67%	71,46%
	300 ms	66,67%	67,50%	65,83%
	600 ms	66,88%	65,00%	58,12%
	1000 ms	62,08%	60,21%	49,58%
FFLFBE_D_Z media, desviación típica	60 ms	60,83%	63,33%	65,83%
	100 ms	60,21%	63,96%	61,88%
	300 ms	62,08%	61,67%	58,96%
	600 ms	63,96%	62,08%	55,21%
	1000 ms	61,88%	57,29%	46,88%
FFLFBE_E_D_Z media, desviación típica	60 ms	61,88%	64,79%	64,38%
	100 ms	61,46%	63,54%	62,08%
	300 ms	63,12%	60,83%	60,21%
	600 ms	63,75%	60,62%	55,21%
	1000 ms	61,88%	58,12%	46,67%

Tabla 21. Resultados de los experimentos por segmento basados en estadísticos a partir de los parámetros FFLFBE para la base de datos SAVEE.

Siguiendo el patrón establecido a partir de los datos obtenidos en los anteriores bloques de experimentos, en la tabla anterior puede observarse cómo el tamaño del segmento tomado tiene una gran influencia sobre los resultados, es decir, a mayor tamaño del segmento peores

resultados. Como en los casos anteriores, la disminución de muestras utilizadas en el entrenamiento del sistema debido al aumento del tamaño de los segmentos produce una disminución en la precisión del sistema.

A diferencia de lo que ocurría en la base de datos EMODB, en la que las mejores tasas de reconocimiento correspondían a una mezcla de 32 gaussianas en la gran mayoría de los casos, en el caso que nos ocupa los mejores valores pertenecen a experimentos en los que se ha llevado a cabo una mezcla de 128 gaussianas.

Si bien es cierto que los mejores valores de este conjunto de experimentos no superan los obtenidos en los de referencia, éstos son ligeramente mejores que los obtenidos siguiendo el mismo proceso en la base de datos EMODB, habiéndose obtenido un 71,40% para un FFLFBE_E_D_Z en el que se han utilizado segmentos de 100 milisegundos y 32 mezclas de gaussianas, mientras que en el caso de la base de datos SAVEE obtenemos como mejor resultado un 72,92% para el caso de un FFLFBE_E_Z con segmentos de 60 milisegundos y una mezcla de 128 gaussianas. Teniendo en cuenta las conclusiones extraídas a partir de los experimentos por trama, podemos confirmar que utilizando los parámetros FFLFBE en esta base de datos, los mejores resultados se obtienen cuando se incluye la energía en el vector de características.

En la siguiente tabla se muestra el conjunto de experimentos en el que se ha añadido el pitch de la señal de audio para comprobar cuál es su efecto en este caso.

		Modelos de 32 mezclas de gaussianas	Modelos de 64 mezclas de gaussianas	Modelos de 128 mezclas de gaussianas	Modelos de 256 mezclas de gaussianas
		Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas	Porcentaje de muestras reconocidas
FFLFBE_Z media y desviación típica	60 ms	63,96%	68,12%	72,08%	
FFLFBE_Z_pitch media y desviación típica	60 ms	65,42%	69,17%	71,46%	70,83%
FFLFBE_Z_pitch_pitchMax_pitchMin media y desviación típica	60 ms	67,71%	70,42%	71,04%	72,50%
FFLFBE_Z media y desviación típica	100 ms	66,04%	69,38%	71,88%	
FFLFBE_Z_pitch media y desviación típica	100 ms	65,62%	70,42%	68,75%	66,88%
FFLFBE_Z_pitch_pitchMax_pitchMin media y desviación típica	100 ms	67,08%	68,54%	71,04%	69,38%
FFLFBE_E_Z media y desviación típica	60 ms	64,17%	67,92%	72,92%	
FFLFBE_E_Z_pitch media y desviación típica	60 ms	66,25%	69,79%	72,29%	73,54%
FFLFBE_E_Z_pitch_pitchMax_pitchMin media y desviación típica	60 ms	67,29%	66,88%	74,58%	72,50%
FFLFBE_E_Z media y desviación típica	100 ms	63,96%	66,67%	71,46%	
FFLFBE_E_Z_pitch media y desviación típica	100 ms	64,58%	68,54%	73,12%	69,17%
FFLFBE_E_Z_pitch_pitchMax_pitchMin media y desviación típica	100 ms	64,38%	67,50%	69,58%	70,21%

Tabla 22. Resultados de los experimentos por segmento basados en estadísticos en presencia y ausencia de la frecuencia fundamental a partir de los parámetros FFLFBE para la base de datos SAVEE.

En contrapartida con los resultados observados en el caso de los coeficientes MFCC, el hecho de añadir el pitch a los vectores de parámetros proporciona un conjunto de valores mejores que los obtenidos para el caso de un experimento basado en estadísticos en el que esto no se ha tenido en cuenta. Por ejemplo, para el caso de un experimento FFLFBE_E_Z en el que se han utilizado segmentos de 60 milisegundos y una mezcla de 128 gaussianas se obtiene un 72,92% de muestras correctamente reconocidas, sin embargo, si para este mismo caso se añade al

vector de características no sólo el pitch sino también su máximo y su mínimo ese valor se convierte en un 74,58%.

Resumen de los resultados de la base de datos SAVEE

Al igual que con la base de datos EMODB, a continuación se muestra una gráfica que contiene un resumen de los mejores resultados en porcentaje, obtenidos a partir del conjunto de muestras acústicas contenidas en la base de datos SAVEE.

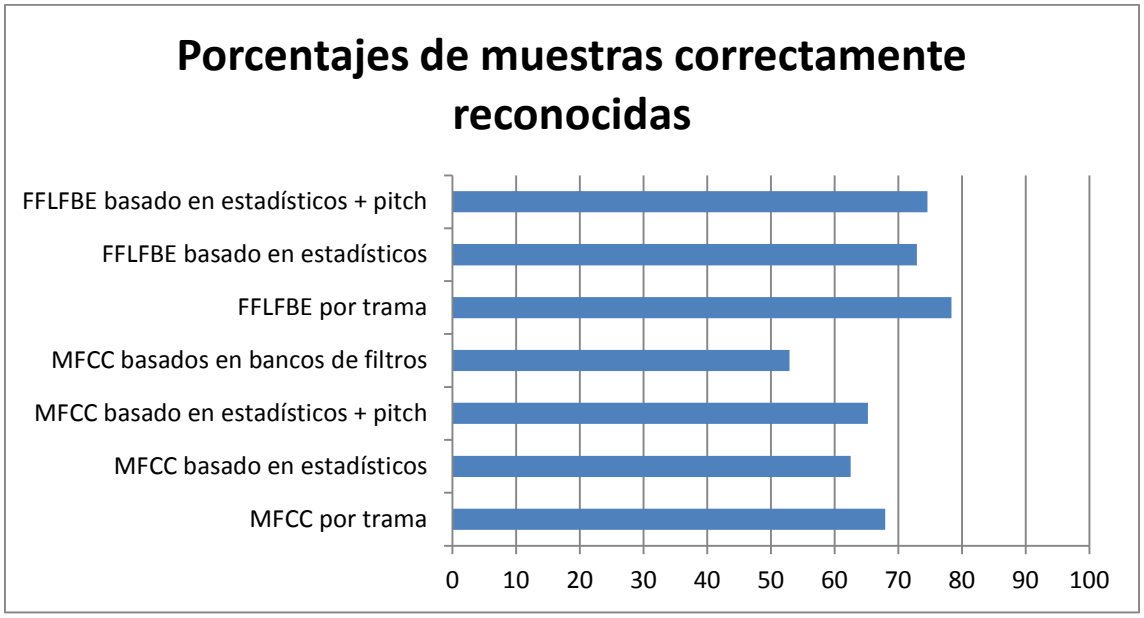

Figura 15. Resumen de los resultados en porcentaje obtenidos para la base de datos SAVEE.

Como puede observarse en la figura anterior, los mejores resultado obtenidos para el conjunto de experimentos realizados en la base de datos SAVEE se corresponden con los experimentos FFLFBE por trama, seguidos de los experimentos FFLFBE basados en estadísticos en los que se ha incluido la frecuencia fundamental y los FFLFBE basados en estadísticos en los que el pitch no está incluido.

6. CONCLUSIONES Y LÍNEAS FUTURAS

A lo largo de este proyecto se ha llevado a cabo un conjunto de experimentos con el fin de extraer una serie de conclusiones acerca de cuáles son los parámetros o características acústicas que proporcionan mejores resultados a la hora de emplear un sistema de reconocimiento de emociones. Con este fin, se han utilizado dos bases de datos con las que se ha pretendido afianzar las conclusiones obtenidas.

En primer lugar, tal y como se ha señalado en el capítulo de Resultados experimentales, dicho conjunto de experimentos se realizó a partir de las muestras contenidas en la base de datos EMODB. En este caso, se llegó a la conclusión de que los experimentos por trama en los que se utilizaban parámetros MFCC proporcionaban los mejores resultados alcanzando un 74,50% (MFCC_E_D_A_Z para una mezcla de 128 gaussianas) frente a un 73,08% obtenido para el caso de los FFLFBE (FFLFBE_E_D_A_Z para una mezcla de 64 gaussianas).

Otra conclusión que pudimos extraer es que al realizar los experimentos basados en estadísticos los resultados obtenidos tienden a empeorar. Esto es debido a que al combinar un determinado número de tramas, para obtener los segmentos que son utilizados en estos experimentos, el número de muestras utilizadas en el entrenamiento del sistema disminuye notablemente. De esta forma, el reconocedor de emociones pierde precisión al tener un menor número de muestras para utilizar en el entrenamiento del mismo y, posteriormente, en el propio reconocimiento. Además, la combinación de dichas tramas para formar estos segmentos, tiene como resultado la pérdida parcial de información contenida en los mismos, pues si bien dichas muestras son independientes del contenido lingüístico de las oraciones utilizadas, la longitud de la misma, en el sentido de cómo se alargan o acentúan las palabras en cuestión, es un parámetro fundamental a la hora de expresar una determinada emoción para un hablante.

Una vez se realizaron estos experimentos a partir de las muestras de la base de datos SAVEE, percibimos diferencias significativas en los resultados obtenidos en estas dos bases de datos. A pesar de conservar una serie de similitudes, como por ejemplo el hecho de que a medida que el tamaño del segmento utilizado aumenta, los resultados obtenidos empeoran, en la base de datos EMODB se observa que los resultados obtenidos en los diversos experimentos nunca superan aquellos que establecimos como referencia. Así mismo, los porcentajes obtenidos a partir de experimentos basados en parámetros MFCC siempre son mejores que aquellos basados en parámetros FFLFBE.

Sin embargo, al repetir estos experimentos utilizando para ello la base de datos SAVEE, es posible comprobar cómo las afirmaciones anteriores no pueden aplicarse a este caso, pues no sólo los parámetros FFLFBE producen mejores resultados que los parámetros MFCC, sino que además dichos resultados pueden llegar a superar, y superan en algunas ocasiones, a los establecidos como referencia.

Otra diferencia entre ambas bases de datos está en la incorporación del pitch o frecuencia fundamental a estos experimentos. Por un lado, en el caso de la base de datos EMODB, la incorporación del pitch al conjunto de parámetros utilizados tanto en el entrenamiento del sistema como en el posterior reconocimiento de emociones produce una disminución de los porcentajes de muestras correctamente reconocidas, llegando a caer de un 72,15% en el caso de un experimento basado en estadísticos MFCC_D_E_Z en el que se ha calculado la media y desviación típica para una mezcla de 32 gaussianas y un segmento de 60 milisegundos a un 67,48% para el mismo caso pero añadiendo al mismo el pitch, o a un 64,11% para el caso en el que además del pitch se ha incluido también la frecuencia fundamental máxima y mínima.

Por otro lado, estos mismos experimentos llevados a cabo en la base de datos SAVEE producen una mejora en los porcentajes de muestras correctamente reconocidas tanto en el caso de los parámetros MFCC como en el de los FFLFBE. Como ejemplo de lo anteriormente expuesto podemos señalar el aumento que se produce en el caso de un experimento basado en estadísticos FFLFBE_E_Z en el que se ha calculado la media y desviación típica para una mezcla de 32 gaussianas y un segmento de 60 milisegundos cuyo porcentaje es 64,17%, a un 66,25% para el mismo experimento pero añadiendo el pitch a este, o a un 67,29% en el caso en el que además de añadir el pitch se ha incluido también la frecuencia fundamental máxima y mínima.

Estas diferencias surgidas por la utilización de distintas bases de datos son debidas a que están grabadas en dos idiomas diferentes por actores cuya lengua materna es dicho lenguaje: alemán en el caso de la base de datos EMODB e inglés en el caso de la base de datos SAVEE. La diferencia de idioma no es el único factor que produce estas diferencias, sino que es posible que también venga unida a una diferencia cultural que produce formas distintas de expresar dichas emociones.

Un hecho que refuerza esta conclusión son los resultados que nos proporcionan los experimentos de referencia. A continuación se muestra un ejemplo que representa lo anteriormente expuesto.

	neut	**ange**	**bore**	**disg**	**fear**	**happ**	**sadn**	**Del [%c / %e]**
neut	51	0	15	11	0	1	1	[64.6/5.2]
ange	1	108	0	6	0	12	0	[85.0/3.6]
bore	22	1	44	5	0	2	7	[54.3/6.9]
disg	9	3	1	22	2	8	1	[47.8/4.5]
fear	10	4	4	6	28	10	7	[40.6/7.7]
happ	1	22	1	5	5	37	0	[52.1/6.4]
sadn	5	0	0	0	0	0	57	[91.9/0.9]

Tabla 23. Modelo de 64 mezclas de un experimento MFCC_E_Z en la base de datos EMODB.

	neut	ange	bore	disg	fear	happ	sadn	Del [%c / %e]
neut	114	0	0	1	0	0	5	[95.0/1.2]
ange	2	37	2	8	2	9	0	[61.7/4.8]
bore	0	3	38	3	9	7	0	[63.3/4.6]
disg	14	4	2	31	6	1	2	[51.7/6.0]
fear	5	3	11	1	32	6	2	[53.3/5.8]
happ	4	6	6	2	7	35	0	[58.3/5.2]
sadn	21	0	0	1	3	1	34	[56.7/5.4]

Tabla 24. Modelo de 64 mezclas de un experimento MFCC_E_Z en la base de datos SAVEE.

Las tablas anteriores muestran un ejemplo de un experimento por trama MFCC_E_Z en cada una de las bases de datos. En la base de datos EMODB, cuyo idioma es el alemán, se puede observar cómo la alegría se confunde de forma muy marcada con el enfado, mientras que en la base de datos SAVEE que está grabada en inglés esta confusión no existe sino que es la tristeza la que se confunde con la neutralidad.

Con este ejemplo se pretende demostrar que un reconocedor de emociones depende del idioma de las muestras utilizadas puesto que en cada idioma estas emociones se expresan de forma diferente en el lenguaje.

Para una futura línea de investigación, realizar esta serie de experimentos en bases de datos de distintos idiomas podría arrojar resultados muy interesantes. Además, la posibilidad de comparar resultados de dos o más bases de datos grabadas en el mismo idioma podría complementar los experimentos anteriores.

7. BIBLIOGRAFÍA

[1] *Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G;* Emotion recognition in human-computer interaction in IEEE Signal Processing Magazine, Vol. 18, Issue 1, pp. 32-80, January 2001.

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

[3] *Dimitrios Ververidis, Constantine Kotropoulos;* Emotional speech recognition: Resources, features, and methods in Elsevier Speech Communications, Vol. 48, Issue 9, pp. 1162-1181, April 2006.

[4] *Björn Schuller,Dejan Arsić, Frank Wallhoff, Gerhard Rigoll;* Emotion recognition in the noise applying large acoustic feature sets in Proc. Speech Prosody 2006, Dresden, Germany, 2006, ISCA, p. no pagination.

[5] *Mohammad Shami, Werner Verhelst;* An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech in Elsevier Speech Communications, Vol. 49, pp. 201-212, January 2007.

[6] *Björn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, Andreas Wendemuth;* Acoustic emotion recognition: a benchmark comparison of performances in Proc. of ASRU 2009, Merano, Italy (2009).

[7] *R. Sahak, Y. K. Lee, W. Mansor, A. I. M. Yassin, A. Zabidi;* Optimized Support Vector Machine for Classifying Infant Cries with Asphyxia using Orthogonal Least Square in 2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2010), December 5-7, 2010, Kuala Lumpur, Malaysia.

[8] *Jeong-Sik Park, Ji-Hwan Kim, Yung-Hwan Oh;* Feature Vector Classification based Speech Emotion Recognition for Service Robots in IEEE Transactions on Consumer Electronics, Vol. 55, No. 3, pp 1590-1596, August 2009.

[9] *Björn Schuller, Anton Batliner, Stefan Steidl, Dino Seppi;* Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge in Elsevier Speech Communications, Vol. 53, pp. 1062-1087, 2011.

[10] *F. Burkhard, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss;* A Database of German Emotional Speech in Proceedings of Interspeech, Lisbon, 2005.

[11] *Philip Jackson and Sanaul Haq;* Surrey Audio-Visual Expressed Emotion (SAVEE) Database in University of Surrey.

[12] *John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue;* TIMIT Acoustic-Phonetic Continuous Speech Corpus in Linguistic Data Consortium, Philadelphia, 1993.

[13] *Beth Logan;* Mel Frequency Cepstral Coefficients for Music Modelling in Cambridge Research Laboratory.

[14] *Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland;* The HTK book for HTK Version 3.2.1.

[15] *Ascensión Gallardo Antolín;* Reconocimiento de habla robusto frente a condiciones de ruido aditivo y convolutivo (Tesis doctoral) en Universidad Politécnica de Madrid, 2002.

[16] *Jerónimo Arenas-García, Jan Larsen, Lars Kai Hansen and Anders Meng;* Optimal filtering of dynamics in short-time features for music organization in 7th International Conference on Music Information Retrieval (ISMIR 2006).

[17] *L. E. Baum and J. A. Eagon;* An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology in Bull. Amer. Math. Soc., 73, 360 – 363 (1967).

[18] *Leonard E. Baum and Ted Petrie;* Statistical Inference for Probabilistic Functions of Finite State Markov Chains in The Annals of Mathematical Statistics, Vol. 37, No. 6 (Dec., 1966), pp. 1554-1563

[19] *Leonard E. Baum, Ted Petrie, George Soules and Norman Weiss;* A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains in The Annals of Mathematical Statistics, Vol. 41, No. 1 (Feb., 1970), pp. 164-171

[20] *L. E. Baum and G. R. Sell;* Growth transformations for functions on manifolds in Pacific Journal of Mathematics, 27, No. 2, 211 – 227 (1968).

[21] *T. Benesch;* The Baum – Welch algorithm for parameter estimation of Gaussian Autoregressive Mixture Models in Journal of Mathematical Sciences, Vol. 105, No. 6, 2001

[22] *G. D. Forney;* The Viterbi algorithm in Proceedings of the IEEE, ISSN 0018-9219, 1973, Vol. 61, No. 3, pp. 268 – 278.