Universidad Autónoma de Madrid
Escuela Politécnica Superior

Proyecto fin de carrera

DETECCIÓN DE EMOCIONES EN VOZ ESPONTÁNEA

Ingeniería Superior en Telecomunicación

Carlos Ortego Resa

Julio 2009

DETECCIÓN DE EMOCIONES EN VOZ ESPONTÁNEA

AUTOR: Carlos Ortego Resa
TUTOR: Ignacio López Moreno

Grupo ATVS
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio 2009

Resumen

Resumen

En este proyecto de fin de carrera se presentan nuevos métodos además del estado del arte de las técnicas existentes para el reconocimiento automático de emoción en el habla. Se emplean técnicas discriminativas como SVM (*Support Vector Machines*) y estadísticas como GMM (*Gaussian Mixture Models*). A partir de estas técnicas se implementan dos tipos de sistemas: *front-end* y *back-end*. Los primeros usan la señal de voz como entrada y producen a la salida unas puntuaciones. Los segundos utilizan como entrada las puntuaciones de salida del sistema *front-end* para obtener a la salida otras puntuaciones.

Se realizará además un examen completo de estos sistemas, desde el conjunto de datos de entrenamiento y test, influencia de distintas variables en los modelos entrenados, fusión de sistemas, normalización de puntuaciones, etc.

En la parte experimental del proyecto se llevan a cabo experimentos independientes y dependientes de locutor con el fin de valorar la variabilidad de locutor sobre los sistemas.

En la memoria se describe el funcionamiento de un sistema automático de reconocimiento de patrones así como los modos de funcionamiento. También se explican los principios básicos de las emociones y cómo afectan éstas al habla. Además, se hace un repaso de las disciplinas más empleadas en el reconocimiento de emociones.

Por último se realiza un análisis del trabajo extrayendo conclusiones y proponiendo futuras líneas de investigación.

Los resultados obtenidos en este proyecto de fin de carrera han sido aceptados y a la espera de ser publicados en 2 congresos internacionales en los artículos:

• Lopez-Moreno, I., Ortego-Resa C., Gonzalez-Rodriguez J., Ramos D. , "Speaker dependent emotion recognition using prosodic supervectors", 2009.

• Ortego-Resa C., Lopez-Moreno, I., Gonzalez-Rodriguez J., Ramos D. , "Anchor model fusion for emotion recognition in speech", 2009.

Palabras Clave

Reconocimiento automático de emociones en el habla, pitch, T-norm, Máquinas de Vectores Soporte, Modelos de Mezcla de Gaussianas, base de datos SUSAS, parametrización prosódica, Fusión de Anchor Models.

Abstract

In this master's thesis we present new methods besides the state of the art of the existing techniques for automatic emotion recognition in speech. Discriminative techniques such as SVM (*Support Vector Machines*) and statistic ones such as GMM (*Gaussian Mixture Models*) are employed. With these techniques two kind of systems are developed: *front-end* and *back-end*. The first one uses voice signal as input signal and a set of scores are obtained as output signal. The second one uses the output scores from *front-end* system as input signal and makes another set of scores as output.

We report a study of these systems regarding training and testing set selection, system behavior acording to some variables, fusion techniques, scores normalizations, etc.

Along the experimental section of the master's thesis several speaker independent and dependent experiments are showing with the purpose of evaluating the speaker variability about systems.

The report describes the operation of an automatic patterns recognition system. It also explains the basic principles of emotions and how they affect speech. In addition, an overview of the disciplines used in emotion recognition is made.

Finally, an analysis of work and conclusions are drawn, and future researchs are proposed.

Results from this master's thesis have been accepted in international congresses and now it is waiting for being published:

• Lopez-Moreno, I., Ortego-Resa C., Gonzalez-Rodriguez J., Ramos D. , "Speaker dependent emotion recognition using prosodic supervectors", 2009.

• Ortego-Resa C., Lopez-Moreno, I., Gonzalez-Rodriguez J., Ramos D. , "Anchor model fusion for emotion recognition in speech", 2009.

Key words

Automatic emotion recognition in speech, pitch, T-norm, Support Vector Machines, Gaussian Mixture Models, SUSAS database, prosodic parametrization, Anchor Models Fusion.

Agradecimientos

Primero dar las gracias a toda la gente que me ha servido de ayuda durante estos últimos años. En especial a mis padres pues ellos son mi modelo a seguir. A mis tios y primos por lo bien que se han portado conmigo. Y a mis abuelos por el apoyo incondicional hacia su nieto.

También me gustaría agradecer a mi tutor Ignacio López Moreno por su apoyo y dedicación en mi proyecto al igual que al resto del grupo ATVS. Además, agradecer a Joaquín González Rodríguez por darme la oportunidad de formar parte del grupo ATVS.

No quiero olvidarme de todos los buenos compañeros que he hecho durante estos 5 años en la EPS: Ángel, Jesús, Jorge, Soci, David, Pablo, Javi,...

Por último, agradecer a mis amigos de toda la vida por lo mucho que me ayudais y me haceis reir.

Carlos Ortego Resa
Julio de 2009

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica superior de la Universidad Autónoma de Madrid. El proyecto ha sido financiado parcialmente por el Ministerio de Ciencia e Innovación a través del proyecto TEC2006-13170-C02-01.

Indice General

Indice de Figuras

Indice de Tablas

1
Introducción

1.1. Motivación

El reconocimiento de emociones a partir de la señal de voz es una disciplina que está ganando interés en la interacción hombre-máquina. Tiene como objetivo identificar automáticamente el estado emocional o físico del ser humano a través de su voz. A los estados emocionales y físicos del locutor se les conoce como aspectos emocionales de la voz y forman parte de los llamados aspectos paralingüísticos del habla. Aunque el estado emocional no altera el contenido lingüístico, éste es un factor importante en la comunicación humana, ya que proporciona más información que la meramente semántica acerca del interlocutor.

Con el progreso de las nuevas tecnologías y la introducción de sistemas interactivos, se ha incrementado enormemente la demanda de interfaces amigables para comunicarse con las máquinas. Existe un amplio rango de aplicaciones en las tecnologías del habla tales como, *call centers*, sistemas inteligentes de automóvil o en la industria del entretenimiento. Por ejemplo, el proyecto *SmartKom* desarrolla un sistema de reserva de entradas que emplea un reconocedor automático del habla siendo capaz de reconocer el nivel de enfado o frustración de un usuario cambiando su respuesta correspondientemente. El reconocimiento automático de emociones en el habla puede ser empleado por terapeutas como una herramienta de diagnóstico en medicina. En psicología, los métodos de reconocimiento de voz emocional pueden hacer frente con la enorme cantidad de datos en tiempo real, obteniendo de forma sistemática las características del habla que transmiten emoción.

El estudio de las características emocionales del habla no tiene como único objetivo el reconocimiento de emociones. Otro de estos objetivos es la síntesis de voz emocional enfocada principalmente para la comunicación de discapacitados. También, tareas como el reconocimiento del habla emocional o el reconocimiento de locutor a partir de voz emocionada son otras de las disciplinas hacia las que está enfocada el estudio de las emociones en el habla.

Por lo general, las emociones no son genuinas o protípicas, sino que se aparecen como combinación de varias. Ésto hace de su clasificación una tarea ardua y dada a la subjetividad. Sin embargo, la mayoría de los investigadores han tratado con emociones prototípicas o completas pues es la única manera de poder discriminar entre unas emociones y otras.

En este proyecto se profundiza en el reconocimiento automático de emociones en el habla. Esta tarea consiste en un problema de clasificación multiclase, donde una locución de habla dada es clasificada entre un conjunto de n emociones. Sin embargo, también resulta de interés detectar una emoción determinada en un segmento de habla, lo cual justifica el uso de un enfoque de verificación o detección.

1.2. Objetivos

El objetivo del presente proyecto es evaluar el funcionamiento de un conjunto técnicas existentes para el reconocimiento de emociones así como de nuevas técnicas presentadas en el mismo. Dichas técnicas se evaluarán tanto para experimentos independientes como dependientes de locutor. En concreto, los sistemas están basados en Modelos de Mezcla de Gaussianas y Máquinas de Vectores Soporte. El proyecto estudia la forma de optimizar los resultados.

Para el entrenamiento de los modelos y para la evaluación de los sistemas, se hace uso de la bases de datos *SUSAS Simulated*, *SUSAS Actual* y *Ahumada III*. El uso de varias bases de datos para la evaluación de nuestros sistemas va a suponer una mayor variabilidad de

emociones y locutores haciendo que los resultados obtenidos sean más realísticos.

1.3. Organización de la Memoria

La memoria consta de los siguientes capítulos:

1. Introducción: motivación y objetivos del proyecto.

2. Sistemas automáticos de reconocimiento de patrones: repasa la estructura y los diferentes tipos de sistemas de reconocimiento de patrones.

3. Estado del arte en Reconocimiento de Emociones: realiza un repaso de las principales técnicas utilizadas para esta disciplina.

4. Diseño y Desarrollo: describe las bases de datos y sistemas empleados para realizar los experimentos.

5. Pruebas y Resultados: describe las pruebas y optimiza los resultados.

6. Conclusiones y trabajo futuro: Evalúa los resultados obtenidos y propone nuevas líneas de investigación y mejora.

2

Sistema automático de reconocimiento de patrones

Un patrón es una colección de descriptores con los cuales representamos los rasgos de una clase. Así, un sistema automático de reconocimiento de patrones es una técnica que mediante el análisis de las características de cierto elemento, asigna una *etiqueta*, que representa a una clase, a un patrón concreto.

Un tipo de sistema de reconocimiento automático de patrones es la biometría o reconocimiento biométrico. Éste, se basa en los rasgos físicos intrínsecos o conductuales para el reconocimiento único de humanos. Estas características o comportamientos humanos forman parte de lo que conocemos como rasgos biométricos.

Otro tipo de sistema de reconocimiento de patrones automático, aunque no propiamente perteneciente al reconocimiento biométricos, es el reconocimiento automático de emociones. Este tipo de sistema se basa en el análisis de las características particulares de las emociones para clasificar unas frente a otras. La percepción del estado anímico humano puede provenir de varios canales, siendo los dos principales las expresiones faciales obtenidas a partir del canal visual y las expresiones léxico-fonéticas provenientes del habla.

2.1. Introducción

2.2. Estructura General

La estructura que sigue un sistema automático de reconocimiento de patrones es generalmente la misma y es la que aparece en la Figura 1. A partir de ahora nos centramos en los sistemas de reconocimiento automático que utilizan la información emocional de la voz como base para la creación de los patrones para dicho reconocimiento.

Desde que la señal de voz emocionada se expone al sistema hasta que el sistema la reconoce, verifica o rechaza existen una serie de etapas intermedias que se pasan a describir.

Por norma general el usuario sólo tiene acceso al sensor, mediante el cual se extraerá la señal de voz. Dicha frontera viene determinada por la línea discontinua de la Figura 1. Los módulos que aparecen con líneas continuas son aquellos que conforman el sistema base de reconocimiento automático, mientras que los módulos con líneas discontinuas son opcionales y se suelen usar como complemento de los primeros.

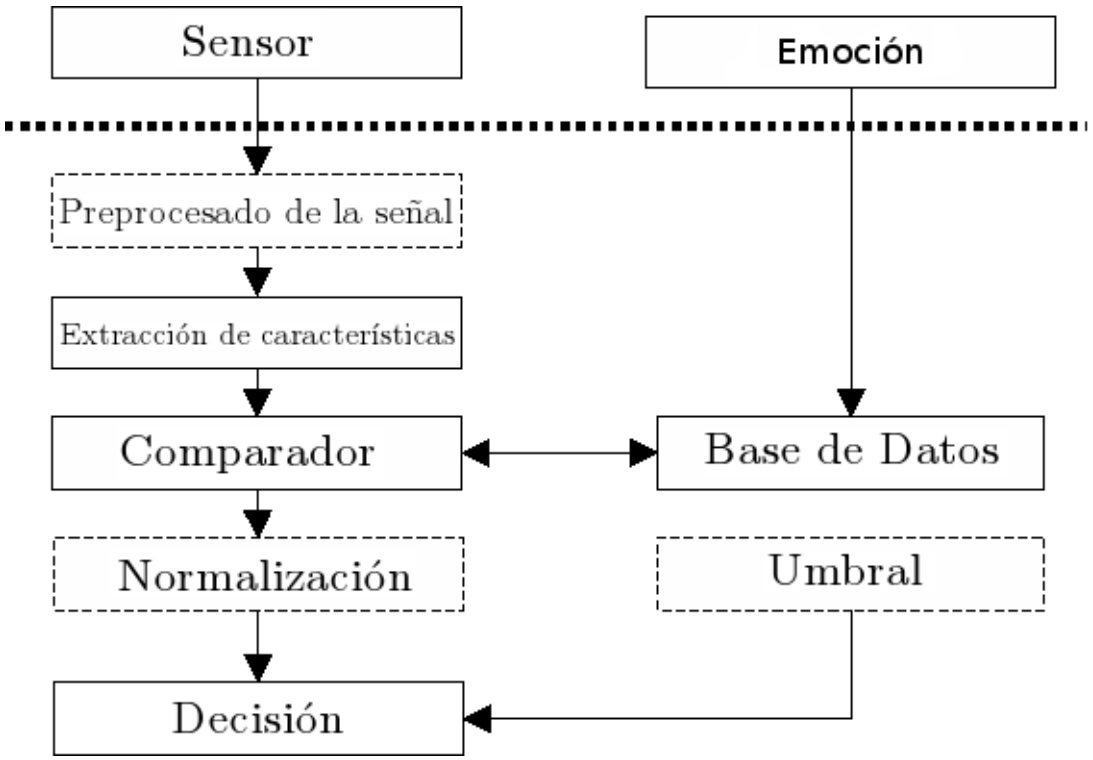

Figura 1: Esquema de funcionamiento de un sistema de reconocimiento.

La primera tarea consiste en la captura por parte de un micrófono de la señal de voz, que transforma la señal acústica en señal eléctrica.

El siguiente paso es la parametrización de la señal de voz o extracción de sus características que consiste en su codificación para que el sistema de reconocimiento sea capaz de medirla y evaluarla cuantitativamente. La parametrización puede venir precedida de un preprocesado de la señal. Esta etapa opcional esta formada por todos aquellas transformaciones que sufre la señal y que facilitan su posterior parametrización o que la hace más eficiente. Un ejemplo de preprocesado es la eliminación de ruido de la señal de voz aplicando diversos tipos de filtros.

Las etapas anteriores son comunes tanto para el proceso de registro como para el de reconocimiento o test. En la etapa de registro, el usuario ofrece su voz al sistema. Ésta es

parametrizada y modelada mediante la fase de entrenamiento para obtener como resultado las diferentes clases (emociones) en que va a poder ser clasificado una muestra de test. Estos modelos se almacenan en una base de datos para la posterior etapa de reconocimiento.

En la etapa de identificación se utiliza un comparador para obtener la similitud de nuestro rasgo parametrizado con respecto a las emociones modeladas en el entrenamiento. Como salida a dicha etapa tenemos una puntuación (*score* en inglés).

La etapa de decisión dependerá del modo de operación del sistema. Si se trabaja en modo de verificación nos hará falta prefijar un umbral que nuestra puntuación ha de sobrepasar para considerar que la emoción de la señal de voz de test pertenece a la emoción objetivo.

2.3. Modos de Operación

Desde el punto de vista de los modos de funcionamiento de los sistemas automáticos de reconocimiento, se puede diferenciar dos perspectivas de trabajo.

- Sistemas de reconocimiento en modo identificación
- Sistemas de reconocimiento en modo verificación

2.3.1. Identificación

El modo de identificación es el que usan los sistemas de reconocimiento automático de locutor e idioma. El objetivo en este tipo de funcionamiento es el de clasificar una realización determinada de un rasgo biométrico como perteneciente a uno de las N posibles clases. Para ello se lleva a cabo una comparación "uno a varios" [Figura 2]. El sistema decidirá si el rasgo de test pertenece a alguna de las clases modeladas en la etapa de entrenamiento o a ninguna. Dentro de estos sistemas debemos de diferenciar entre dos posibles casos.

• **Identificación en conjunto cerrado:** en este caso, el resultado del proceso es una asignación a una de las clases modeladas por el sistema. Existen, por lo tanto, N posibles decisiones de salida posibles.

• **Identificación en conjunto abierto:** aquí debemos de considerar una posibilidad adicional a las N del caso anterior: que el rasgo que pretende ser identificado no pertenezca al grupo de clases que contiene la base de datos, con lo que el sistema de identificación debe de contemplar la posibilidad de no clasificar la realización de entrada como perteneciente a las N posibles.

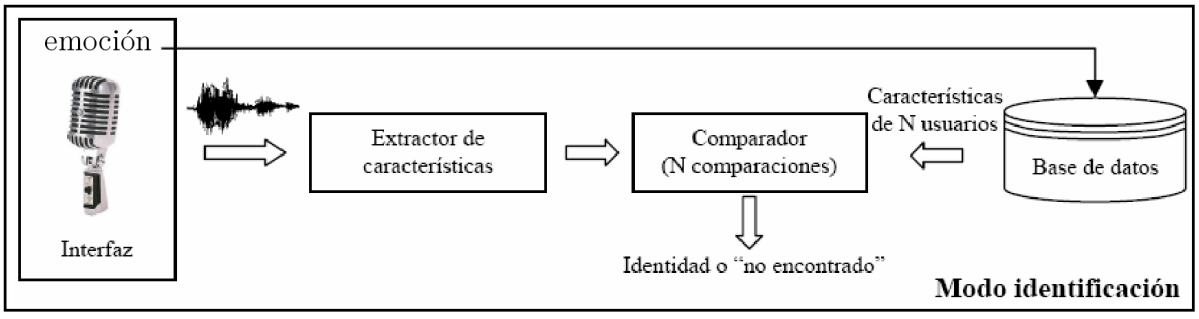

Figura 2: Sistema de reconocimiento automático en modo de identificación. Figura adaptada de [1].

2.3.2. Verificación

Los sistemas de verificación, por el contrario llevan a cabo comparaciones "uno a uno" y por ello suponen un menor coste computacional que el sistema de identificación. [Ver Figura 3]. Este tipo de sistemas necesitan dos entradas: una realización del rasgo de test y una solicitud de identidad a verificar. El sistema busca en la base de datos el modelo de dicha identidad para enfrentarlo a la realización de test facilitada.

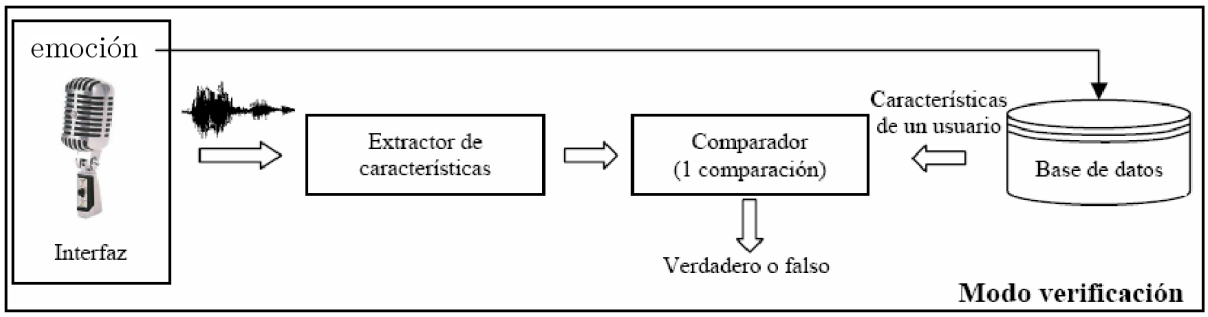

Figura 3: Sistema de reconocimiento automático en modo de verificación. Figura adaptada de [1].

De este modo las dos únicas salidas o decisiones posibles del sistema son la aceptación o rechazo del rasgo de test como aquel que pretende ser. La decisión de aceptación o rechazo dependerá de si la puntuación obtenida en la identificación supera o no un determinado umbral de decisión.

Los sistemas de verificación pueden ser vistos como un caso particular de identificación en conjunto abierto, en el que N=1.

2.4. Evaluación de los Sistemas Automáticos de Reconocimiento

Una de las tareas más importantes a la hora de diseñar un sistema de reconocimiento automático es obtener una medida fiable y precisa de su rendimiento. Gracias a ello vamos a poder determinar si nuestro sistema cumple unas especificaciones mínimas de funcionamiento, evaluar posibles mejoras o compararlo con otros sistemas.

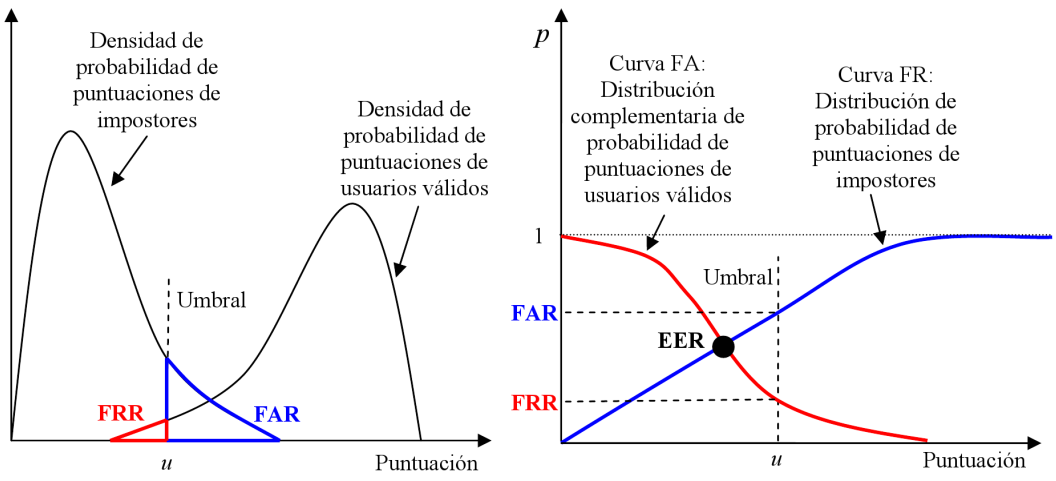

Figura 4: Densidades y distribuciones de probabilidad de intentos target y non-target.

En primer lugar hay que tener claro que son pruebas target y pruebas non-target. Se

denomina intento target cuando se comparara una muestra de una clase con el patrón de la misma clase, si la muestra y el patrón son de clases distintas, al intento se le denomina non-target. Cuanto mayor sea el número de intentos de tanto pruebas target como non-target, más fiable será la medida del rendimiento del sistema. Las puntuaciones obtenidas en pruebas target serán puntuaciones de usuarios válidos mientras que las obtenidas en pruebas non-target serán puntuaciones de usuarios impostores. El comportamiento del sistema dependerá del valor de umbral a partir del cual aceptaré la muestra de prueba como perteneciente a la clase de referencia.

Podemos tener dos tipos de errores, bien que una muestra auténtica sea rechazada, lo que llamaremos tasa de Falso Rechazo (FR), o que una muestra falsa sea aceptada, lo que llamaremos tasa de Falsa Aceptación (FA).

El umbral es un valor que influye directamente en la tasa de falsa aceptación y falso rechazo. Según se puede ver en la Figura 4, un valor alto de umbral hace que pocas pruebas non-target sean aceptadas y por lo tanto la FA descenderá, a costa de incrementar la FR. Por el contrario, un valor pequeño de umbral hace que aumente la FA manteniendo bajo la FR. Como vemos, existe un compromiso entre FR y FA que se debe evaluar acorde a la aplicación a la que vaya dirigido nuestro sistema. Así, por ejemplo, en un control de acceso de alta seguridad sería adecuado trabajar con un elevado valor de umbral impidiendo de este modo una tasa alta de FA.

Se considera el valor de error igual, EER (Equal Error Rate), a aquel punto donde las curvas de falsa aceptación y falso rechazo se cruzan. Esta tasa se suele usar para comparar sistemas.

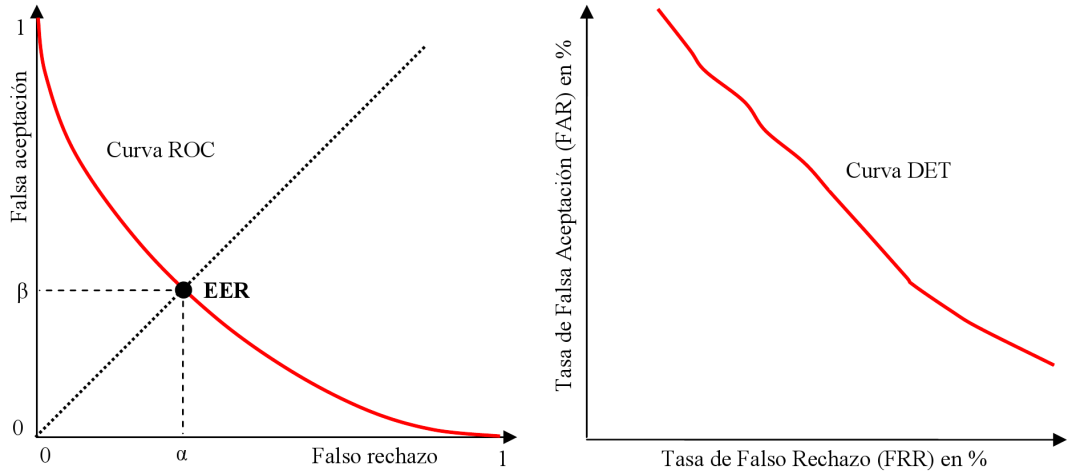

Figura 5: Curvas ROC y DET.

Otro forma de representar gráficamente el rendimiento del sistema es mediante las curvas de la Figura 5. En ellas se enfrenta la probabilidad de FA y FR en una gráfica. Así, podemos ver que valores de probabilidad de FA y FR tenemos para cada umbral escogido. A esta curva se le llama curva ROC (Receiver Operating Curve). Otra alternativa son las curvas DET (Detection Error Tradeoff), cuya única diferencia con las curvas ROC es un cambio de escala en los ejes [3]. Serán las curvas DET las que se usarán en la sección de experimentos para mostrar los resultados de forma gráfica.

Junto a cada una de estas curvas se incluirá una tabla con tres valores importantes a la hora de evaluar un sistema. Estos valores serán: el DCF mínimo (*Detection Cost Function*), EER global (en %) y EER medio (en %).

El EER medio se calcula como el valor medio de los EERs por modelo. Así, EER medio diferirá del EER global cuando los modelos no tengan todos el mismo número de intentos o puntuaciones.

La función de coste es otra forma habitual de medir el rendimiento de los sistemas. Se define como:
$C_{DET}(\mathrm{i}) = C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target} + C_{FalseAlarm} \cdot P_{FalseAlarm|NonTarget} \cdot (1 - P_{Target})$

Donde C_{Miss} es el coste asociado a un falso rechazo, $C_{FalseAlarm}$ es el coste asociado a una falsa aceptación, P_{Target} es la probabilidad de que un fichero dado pertenezca a la emoción en cuestión (establecida a priori), $P_{Miss|Target}$ es el porcentaje de falsos rechazos (dado por el sistema) y $P_{FalseAlarm|NonTarget}$ es la probabilidad de una falsa aceptación (dada por el sistema).

Los valores fijados para la evaluación de locutor NIST SRE 2006 son: C_{miss}=1, $\mathrm{C}_{FalseAlarm}$=10 y P_{Target}=0.01. Estos valores son los que vamos a utilizar para nuestro trabajo pues se ajustan también a nuestra tarea de evaluación de reconocimiento de emociones. Una falsa aceptación se penaliza 10 veces más que un falso rechazo. La probabilidad de que el locutor experimenta una cierta emoción es de 0.01 pues lo habitual es encontrarnos en un estado de excitación normal.

De esta forma, con el porcentaje de falsa aceptación y falso rechazo, $\mathrm{P}_{Miss|Target}$ y $\mathrm{P}_{FalseAlarm|NonTarget}$, obtenido de nuestro sistema podremos evaluar la función de coste, obteniendo lo que se conoce como DCF. El DCF mínimo será el mínimo valor de la DCF.

2.5. Normalización de Puntuaciones

Los sistemas de reconocimiento automático de patrones producen como salida una serie de puntuaciones que evalúan la similitud entre las muestras de test y las clases o modelos.

Mediante las normalizaciones lo que se pretende es proyectar las puntuaciones tanto de pruebas target como non-target sobre un espacio acotado de media cero y varianza unidad, de tal modo que las puntuaciones queden acotadas.

Con dicha proyección o escalado de las puntuaciones, lo que se pretende es buscar un umbral global para la tarea de decisión ya que puede ocurrir que durante la fase de autenticación, las puntuaciones de un determinado usuario, tanto las del propio usuarios como las de los impostores, estén en un rango de valores distinto al de otros usuarios. Este efecto se conoce como desalineamiento. La normalización de puntuaciones son el conjunto de técnicas y algoritmos que permiten aumentar el rendimiento y robustez del sistemas compensando este desalineamiento.

La normalización de puntuaciones es también importante para la posterior fusión de sistemas pues sitúa las puntuaciones de sistemas individuales en rangos homólogos.

Las técnicas más corrientes de normalización de puntuaciones son la T-norm y la Z-norm. La T-norm (Test Normalization) [4] es una normalización dependiente de la muestra de test, mientras que la Z-norm (Zero Normalization) [4] es dependiente del modelo o usuario.

3

Estado del arte en Reconocimiento de Emociones

El reconocimiento automático de emociones es sin duda una tarea multidisciplinar que involucra diferentes campos de investigación tales como psicología, lingüística, análisis de voz, análisis de imágenes y aprendizaje automático. El progreso en el reconocimiento automático de emociones está condicionado al progreso en cada uno de los campos.

Por ello, un sistema reconocedor debería de realizar un análisis multimodal en el cual interviniese información procedente de diferentes sensores. Hay muchas señales humanas a partir de las cuales se puede sacar información sobre el estado emocional de la persona, como por ejemplo, la voz, la imagen facial, gestos y posturas, ritmo de respiración y latido del corazón, etc. Las tareas más estudiadas actualmente son el reconocimiento de emociones en el habla y en imágenes faciales. Si bien este proyecto sólo analizará el reconocimiento de emociones en el habla.

3.1. Introducción

En la comunicación humana se puede distinguir dos canales diferenciados. Uno de ellos se encarga de transmitir el mensaje de forma explícita, es decir, el contenido meramente semántico. El otro tipo de canal no explícito hace enriquecer la comunicación humano-humano y es el que transmite información implícita como edad, sexo, estado emocional del usuario, etc. Es en éste en el que se centra el reconocimiento automático de emociones. La importancia de estudiar el reconocimiento emocional y añadirlo a una interfaz automática es grande ya que es la base de las relaciones humanas, y se fundamenta en la interpretación de las señales transmitidas de forma inconsciente y que no siempre son verbales.

El paradigma de la comunicación hombre-máquina sugiere que las interfaces futuras se deben centrar en el humano y ser capaces de anticiparse, como por ejemplo, teniendo la habilidad de detectar cambios en el comportamiento del usuario, especialmente su comportamiento emocional.

3.2. Aplicaciones

Los sistemas de reconocimiento automático de emociones están orientados hacia una amplia gama de aplicaciones. Se podría diferenciar entre dos tipos de campos de aplicaciones; aquellas que mejoran la calidad de vida, y las que sirven para mejorar investigaciones relacionadas con la emoción [5].

Entre las aplicaciones cuya finalidad es mejorar la calidad de vida tenemos servicios al cliente sensibles a la emoción, *call centers*, sistemas de automóviles inteligente capaces de detectar fatiga en el conductor, aplicaciones orientadas a la industria del juego y entretenimiento o sistemas de síntesis de habla emocional para discapacitados. Estos sistemas cambiarán la manera en que interaccionamos con las máquinas. Por ejemplo, un servicio de *call center* automático con detector de emoción sería capaz de producir una respuesta apropiada o pasar el control a un operador humano. La mayoría de los sintetizadores de habla actuales ofrecen voz neutra que resulta monótona y rutinaria. El proveer a estos sistemas de voz personalizada sería de gran ayuda para personas disminuidas.

El otro grupo importante de aplicaciones está orientado a la mejora de investigaciones (por ejemplo, en psicología, psiquiatría, comportamiento humano o neurología), donde este tipo de sistemas puede mejorar la calidad de la investigación obteniendo mayor fiabilidad en las medidas y mayor velocidad en tareas manuales de procesado de datos sobre el comportamiento emocional. Las áreas de investigación en las que se puede obtener un beneficio sustancial son investigaciones en la conducta social (como el grado de interés de un sujeto en la comunicación [6]) y emocional, la relación madre-hijo, trastornos psiquiátricos y el estudio de expresiones afectivas (por ejemplo, decepción).

3.3. Naturaleza de las Emociones

En cada instante experimentamos algún tipo de emoción o sentimiento. Nuestro estado emocional varía a lo largo del día en función de lo que nos ocurre y de los estímulos que percibimos. Otra cosa es que tengamos siempre conciencia de ello, es decir, que sepamos y podamos expresar con claridad que emoción experimentamos en un momento dado.

Las emociones son experiencias muy complejas y para expresarlas utilizamos una gran variedad de términos, además de gestos y actitudes. Debido a su complejidad sería imposible

hacer una descripción y clasificación de todas las emociones que podemos experimentar. Sin embargo, el vocabulario usual para describir las emociones es reducido y ello permite que las personas de un mismo entorno cultural puedan compartirlas.

La complejidad con la que podemos expresar nuestras emociones nos hace pensar que la emoción es un proceso multifactorial o multidimensional. Uno siempre tiene la impresión de que le faltan palabras para describir con precisión sus emociones. La emoción no es un fenómeno simple, sino que muchos factores contribuyen a ello. Se experimentan a veces cuando algo inesperado sucede y los efectos emocionales empiezan a tener control en esos momentos.

Emoción y estado emocional son conceptos diferentes: mientras que las emociones surgen repentinamente en respuesta a un determinado estímulo y duran unos segundos o minutos, los estados de ánimo son más ambiguos en su naturaleza, perdurando durante horas o días. Las emociones pueden ser consideradas más claramente como algo cambiante y los estados de ánimo son más estables. Aunque el principio de una emoción puede ser fácilmente distinguible de un estado de ánimo, es imposible definir cuando una emoción se convierte en un estado de ánimo; posiblemente por esta razón, el concepto de emoción es usado como un término general que incluye al del estado de ánimo.

Como término más general al de estado de ánimo y emoción, está el rasgo a largo plazo de personalidad, que puede definirse como el tono emocional característico de una persona a lo largo del tiempo.

Muchos de los términos utilizados para describir emociones y sus efectos son necesariamente difusos y no están claramente definidos. Esto es atribuible a la dificultad en expresar en palabras los conceptos abstractos de los sentimientos, que no pueden ser cuantificados. Por ello, para describir características de las emociones se utilizan un conjunto de palabras emotivas, siendo seleccionadas la mayoría de ellas por elección personal en vez de comunicar un significado estándar.

3.4. Emociones en el Habla

La voz es el principal modo de comunicación entre humanos y por consiguiente a lo largo de las últimas décadas se ha estudiado las maneras en que funciona el tracto vocal a la hora de producir voz. Durante este tiempo se ha investigado la manera de diseñar sistemas capaces de sintetizar y reconocer voz electrónicamente.

Uno de los mayores problemas con los que se ha encontrado la comunidad científica a la hora de estudiar los mecanismos del habla es la variabilidad de ésta. Muchos estudios han demostrado que por medio de la voz se es capaz de reconocer varios aspectos del estado físico, tales como la edad, sexo, apariencia y del estado emocional [7], [8]. Todo este conjunto de factores, diferentes para cada locutor, contribuyen a la variabilidad del habla. El problema por ejemplo en los sintetizadores de habla es que no ofrecen esta variabilidad en el habla y producen por lo tanto un habla no natural. La variabilidad en el habla supone también un problema en el reconocimiento de habla haciendo así que un contenido semántico como por ejemplo una palabra pueda ser expresada de un número incalculable de maneras dependiendo de las condiciones de cada locutor, sexo, edad, estado emocional, etc.

Para implementar con éxito los reconocedores de emociones en el habla hay que tener en cuenta dos factores fundamentales: el conocimiento de como pueden distinguirse las características emocionales de la voz y como pueden describirse dichas características usando los métodos de procesado de voz convencionales.

Si consideramos el conjunto de características del habla que puedan ser analizadas en habla emocionada (bajo estrés), la frecuencia fundamental o pitch es una de las que más se ha estudiado históricamente. Uno de los primeros y más amplios trabajos sobre el análisis de las características del habla fue Williams y Stevens [9], al cual le fueron sucediendo más con el tiempo.

Los efectos de las emociones en el habla han sido estudiados por investigadores acústicos que han analizado la señal de voz, por lingüistas que han estudiado los efectos léxicos y prosódicos y por psicólogos. Gracias a estos esfuerzos se ha conseguido identificar muchos de los componentes del habla que se utilizan para expresar emociones, dentro de los cuales se consideran los más importantes: pitch, duración, calidad de voz y forma del pulso glotal y tracto vocal.

3.4.1. Pitch

El pitch o frecuencia fundamental es la frecuencia a la que vibran las cuerdas vocales, también llamada frecuencia fundamental o F0. Es uno de los parámetros que caracterizan la voz de un locutor. Se considera que las características del pitch son unas de las principales portadoras de la información emocional.

Las características de la frecuencia fundamental incluyen contorno, media, variabilidad y distribución.

• El valor medio del pitch depende del locutor y expresa el nivel de excitación del locutor. Podemos afirmar que una media elevada de F0 indica un mayor grado de excitación.

• El rango del pitch es la distancia entre el valor máximo y mínimo de la frecuencia fundamental. Refleja también el grado de exaltación del locutor. Un rango más extenso que el normal refleja una excitación emocional o psicológica.

• Las fluctuaciones en el pitch descritas como la velocidad de la fluctuaciones entre valores altos y bajos y si son abruptas o suaves son producidas psicológicamente. En general, la curva de tono es discontinua para las emociones consideradas como negativas (miedo, enfado) y es suave para las emociones positivas (por ejemplo la alegría).

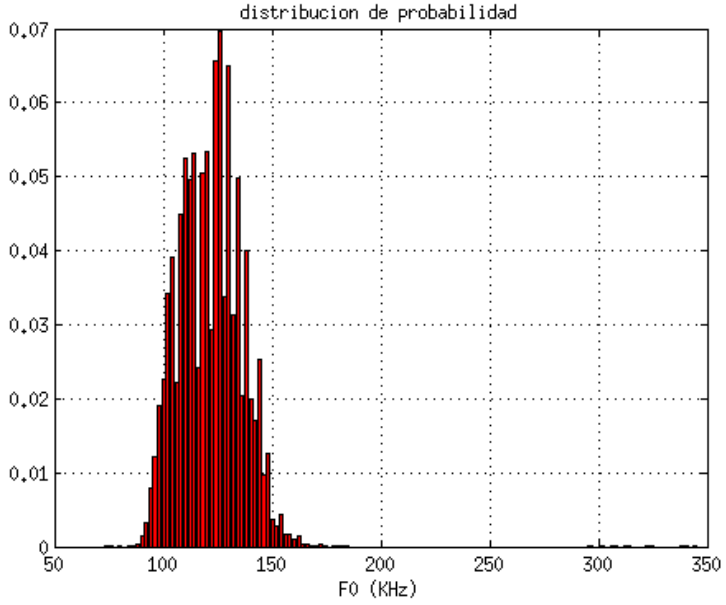

Figura 6: Ejemplo de distribución de probabilidad de pitch para un locutor masculino.

• La distribución de pitch describe el rango de valores de pitch así como la probabilidad de que un cierto valor esté dentro de un subconjunto de dicho rango. Dicha distribución es precisamente lo que modelaremos posteriormente en nuestro sistema GMM-SVM. La Figura 6 corresponde con un ejemplo de distribución de pitch de un locutor masculino. Una de las maneras más fácil de distinguir entre voz masculina y femenina es a través de la distribución del pitch. Así, el género femenino posee una frecuencia fundamental media aproximadamente el doble a la del hombre y una desviación también mucho mayor [Figura 7], es decir, existe mayor diversidad de tono de voz en mujeres que en hombres.

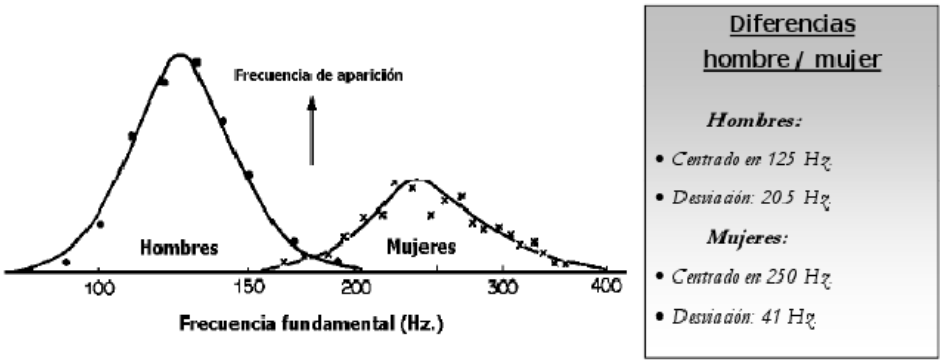

Figura 7: Distribución F0 hombre/mujer.

3.4.2. Duración

La duración es la componente de la prosodia descrita por la velocidad del habla y la situación de los acentos, y cuyos efectos son el ritmo y la velocidad. El ritmo en el habla deriva de la situación de los acentos y de la combinación de las duraciones de las pausas y de los fonemas.

Para ciertas condiciones de estrés, la duración de las palabras o de los fonemas, los cambios entre vocales frente a consonantes o la presencia de consonantes juegan un papel importante en la habilidad de los oyentes a la hora de recibir la información del locutor [10].

Las emociones pueden distinguirse por una serie de parámetros que conciernen a la duración, como son:

• velocidad de locución: generalmente un locutor en estado de excitación acortará la duración de las sílabas, con lo que la velocidad de locución medida en sílabas por segundo o en palabras por minuto se incrementará.

• número de pausas y su duración: un locutor exaltado tenderá a hablar rápidamente con menos pausas y más cortas, mientras que un locutor deprimido hablará más lentamente, introduciendo pausas más largas.

3.4.3. Calidad de Voz

La calidad de voz puede marcar la diferencia entre unas emociones y otras. Existen numerosas variables fonéticas relacionadas con la calidad de voz: cociente de abertura de las cuerdas vocales, timbre e irregularidades de la voz, ruido, distribución de la energía (intensidad), laringerización, etc.

3.4.4. Pulso Glotal y Tracto Vocal

Las características espectrales producidas como respuesta al tracto vocal y glotal también se ven modificadas durante la producción de habla bajo estrés.

Características de la forma del pulso glotal como la pendiente, centro de masas o nivel medio espectral, también han sido analizadas como potenciales rasgos acústicos correlados con el habla emocional. También han sido investigadas la media, varianza y la localización y ancho de banda de los formantes para estudiar el efecto del habla bajo condiciones de estrés [11].

La Tabla 1 presenta un resumen de las relaciones entre las emociones y los parámetros del discurso. Como se puede observar en la tabla únicamente aparecen cinco emociones. Como veremos en la sección 3.5, éstas corresponden con las emociones primarias o básicas. El resto de emociones modifican y combinan estas emociones básicas y son las que llamamos emociones secundarias.

	Ira	***Felicidad***	***Tristeza***	***Miedo***	***Disgusto***
Veloc. Habla	Ligeramente acelerada	Acelerada o retardada	Pausada	Muy acelerada	Mucho más acelerada
Calidad voz	Procedente del pecho	Estridente	Resonante	Irregular	Retumbante
Intensidad	Alta	Alta	Baja	Normal	Baja
Pulso glotal	Pendiente fuerte y alto ancho banda	Pendiente fuerte	Pendiente suave y ancho banda estrecho.	Pendiente muy fuerte y gran ancho banda	Pendiente fuerte

Tabla 1: Emociones y características del habla.

Existe en general una relación conocida entre el habla y las emociones primarias. Las medidas del habla que parecen ser buenas indicadoras de estas emociones son medidas acústicas continuas, tales como las relacionadas con la variación del discurso, el rango, la intensidad y la duración del mismo. Sin embargo esta relación suele no ser suficiente. Una de las líneas de investigación en el reconocimiento automático de emociones es la mejora de nuestra capacidad para identificar la correlación entre las señales acústicas en el discurso y el amplio rango de emociones producidas por el hablante. Los sistemas diseñados para llevar a cabo esta tarea, por lo general, son extremadamente sensibles a la variabilidad introducida por el hablante. Esta variabilidad se debe, especialmente a variaciones en la voz y en estilo causadas por ejemplo por diferentes estados de ánimo del hablante [12].

3.5. Clasificación de las Emociones

En la mayoría de los casos, las emociones no son genuinas o protípicas, sino que se dan como mezcla de varias. Ésto provoca que la clasificación de las emociones sea una tarea ardua y totalmente expuesta a las subjetividad. Sin embargo, la mayoría de los investigadores han tratado siempre con emociones prototípicas o completas pues es la única manera de poder discriminar entre unas emociones y otras.

Basándonos en el grado en que las emociones afectan al comportamiento del sujeto podemos clasificar las emociones como positivas o negativas. Cada emoción expresa una cantidad o

magnitud en una escala positivo/negativo. Así, experimentamos emociones positivas y negativas en grados variables y de intensidad diversa. Podemos experimentar cambios de intensidad emocional bruscos o graduales, bien hacia lo positivo o bien hacia lo negativo. Es decir, toda emoción representa una magnitud o medida a lo largo de un continuo, que puede tomar valores positivos o negativos.

En el lenguaje cotidiano, expresamos nuestras emociones dentro de una escala positivo-negativo y en magnitudes variables, como "me siento bien", "me siento muy bien", "me siento extraordinariamente bien"(intensidades o grados del polo positivo) o "me siento mal", "me siento muy mal", "me siento extraordinariamente mal"(intensidades o grados del polo negativo).

Según sea la situación que provoca la emoción, escogemos unas palabras u otras como 'amor', 'amistad', 'temor', 'incertidumbre', 'respeto', etc., que, además, señala su signo (positivo o negativo). Y según sea la intensidad de la emoción escogemos palabras como 'nada', 'poco', ' bastante', 'muy', etc. y así, componemos la descripción de una emoción. Decimos, por ejemplo, "me siento muy comprendido"(positiva) o "me siento un poco defraudado"(negativa).

En consecuencia, podemos reconocer en toda emoción dos componentes bien diferenciados. Por un lado, un componente cualitativo que se expresa mediante la palabra que utilizamos para describir la emoción (amor, amistad, temor, inseguridad, etc.) y que determina su signo positivo o negativo. Por otro lado, toda emoción posee un componente cuantitativo que se expresa mediante palabras de magnitud (poco, bastante, mucho, gran, algo, etc.), tanto para las emociones positivas como negativas.

Otro tipo de clasificación es la que diferencia entre emociones primarias y emociones secundarias. Las primeras son las emociones fundamentales mientras que las secundarias son todas las demás que modifican y combinan estas emociones básicas. Sin embargo, no hay consenso sobre cuáles constituyen las emociones básicas.

- **Emociones primarias**

- Enfado: El enfado ha sido ampliamente estudiado en la literatura sobre emociones. Hay contradicciones entre los efectos recogidos en estos escritos, aunque esto puede ser debido a que el enfado puede ser expresado de varias maneras. El enfado se define como "la impresión desagradable y molesta que se produce en el ánimo". El enfado se caracteriza por un tono medio alto (229 Hz), un amplio rango de tono y una velocidad de locución rápida (190 palabras por minuto), con un 32 % de pausas.

- Alegría: Se manifiesta en un incremento en el tono medio y en su rango, así como un incremento en la velocidad de locución y en la intensidad.

- Tristeza: El habla triste exhibe un tono medio más bajo que el normal, un estrecho rango y una velocidad de locución lenta.

- Miedo: Comparando el tono medio con los otras cuatros emociones primarias estudiadas, se observó el tono medio más elevado (254Hz), el rango mayor, un gran número de cambios en la curva del tono y una velocidad de locución rápida (202 palabras por minuto).

- Disgusto/odio: Se caracteriza por un tono medio bajo, un rango amplio y la velocidad de locución más baja, con grandes pausas.

- **Emociones secundarias**

- Pena: es una forma extrema de tristeza, generalmente causada por una aflicción. Se

caracteriza por un bajo tono medio, el rango de tono más estrecho, la pendiente de la curva de tono más baja, una velocidad de locución baja y un alto porcentaje de pausas.

- Ternura: se expresa con un alto nivel de tono que no fluctúa excesivamente.

- Ironía: caracterizada por una velocidad de locución baja y una acentuación muy marcada.

- Sorpresa: con un tono medio mayor que la voz normal, una velocidad igual a la normal y un rango amplio.

Otras emociones secundarias: como el temor, la queja, el anhelo, el aburrimiento, la satisfacción, la impaciencia, el ensueño, la coquetería han sido también objeto de estudio.

Autores como J. Davitz, Osgood, Suci y Tannembaum clasificaron las emociones utilizando para ello tres dimensiones del espacio semántico: potencia, valencia y actividad [13].

• **Potencia o fuerza:** corresponde a la atención – rechazo. Ayuda a distinguir entre emociones iniciadas por el sujeto a aquellas que surgen del ambiente (desde el desprecio al temor o la sorpresa). También se le ha llamado fuerza o dominio.

• **Valencia, agrado o valoración:** se refiere al grado de positividad o negatividad de la emoción (desde la alegría hasta el enfado).

• **Actividad:** corresponde al grado de intensidad en la emoción. También se la conoce como intensidad o dimensión de intensidad.

En varios estudios se ha descubierto que se confunden más entre sí las emociones con un nivel similar de actividad (como por ejemplo la alegría y el enfado) que las que presentan similitud en términos de valencia o de fuerza. También están relacionados el ritmo y la valencia de forma que los sentimientos “positivos” son expresados con un ritmo más regular que los sentimientos “negativos”. Esto lleva a la conclusión que la dimensión de la actividad está más correlacionada con las variables auditivas relativamente más simples de la voz, como pueden ser el tono y la intensidad, mientras que la valencia y la fuerza son probablemente comunicados por modelos más sutiles y complejos.

Algunos investigadores han utilizado otra clasificación, dividiendo las emociones en:

• **Pasivas:** Se caracterizan por una velocidad de locución lenta, un volumen bajo, un tono bajo y un timbre más resonante.

• **Activas:** Caracterizadas por una velocidad de locución rápida, alto volumen, alto tono y un timbre ”encendido”.

3.6. Implicaciones Jurídicas

Existen varias áreas donde el reconocimiento de emociones puede influir en una sentencia legal. En lineas generales, estas áreas incluyen valoración de emociones en los demás, emociones y memoria (credibilidad de testigos), emociones y cultura (efectos en investigaciones forenses), y conocimiento legal y emociones [14].

•**Valoración de emociones en los demás**

La capacidad de detectar emociones y el grado de las mismas a través de rasgos acústicos de la señal de voz puede ser de gran utilidad en el sistema jurídico. Por ejemplo, las fuerzas de la ley se pueden beneficiar conociendo que emociones experimenta un sospechoso en un interrogatorio para así evaluar su credibilidad. O un jurado puede dar credibilidad o no a un testigo bajo el conocimiento de cambios en su habla.

•Emociones y memoria

Erróneamente, los tribunales depositan demasiada confianza en los testigos visuales o auditivos. Para calcular de forma más acertada la fiabilidad de los testimonios de los testigos, éstos deberían deberían ser analizados computacional, como por ejemplo, incluir una valoración del estado emocional del testigo. Se necesita de un mejor entendimiento de las emociones pues éstas juegan un papel crucial en la memoria. Los psicólogos cognitivos suelen distinguir entre formación, codificado, asociación y reconstrucción de la memoria. Todos estos procesos pueden ser afectados por la emoción. Se piensa que por ejemplo, los sucesos emocionales conllevan alguna preferencia en su procesado y por lo tanto son más estables y los recordamos con mayor precisión.

•Emociones y cultura

Las diferencias entre las emociones entre culturas puede suponer un serio problema en las investigaciones forenses. Por ejemplo, se ha visto que las interpretaciones de una lengua foránea en interrogatorios policiales generan problemas, especialmente si el intérprete no ha sido entrenado correctamente o si el policía actúa como intérprete. Las traducciones literales de lenguas extrajeras deberían ser enfatizadas para dar un entendimiento global de lo que se quiere comunicar. Sin embargo, no puede ser del todo posible debido a la ambigüedad entre el gran número de traducciones hay entre idiomas y culturas.

•Emociones y conocimiento legal

El sistema judicial reconoce a las emociones como una parte íntegra del mismo. El propio sistema está basado en normas morales, las cuales, se basan en valores emocionales. Por ejemplo, los crímenes se castigan, además de por su carácter intrínseco, por la actitud del culpable sobre la víctima. Así, el castigo se gradúa por las emociones que el culpable padece en los momentos que rodean al acontecimiento. Por lo tanto, las emociones se entrelazan intrínsecamente con la ley.

3.7. Técnicas de Reconocimiento de Emociones

En esta sección se van a mostrar las técnicas de reconocimiento de emociones en el habla más importantes que se estudian en la actualidad. La mayoría de las técnicas usadas ahora para el reconocimiento de emociones anteriormente vienen de ofrecer buenos resultados en tareas de reconocimiento de tanto locutor como de idioma.

3.7.1. GMM

La técnica de Modelos de Mezcla de Gaussianas (*GMM* o *Gaussian Mixture Models*) aplicada al reconocimiento automático de emociones [15] se basa en el principio de que las emociones tienen diferentes sonidos y que la frecuencia de aparición de los sonidos es diferente de una emoción a otra. Los GMM modelan la distribución de probabilidad de los parámetros ($\vec{x}$) de un fragmento de audio. Los parámetros que más se usan son los MFCC (*Mel Frequency Cepstral Coefficients*) o SDC (*Shifted Delta Cepstral*) como parámetros acústicos y contornos

de energía y pitch para parámetros prosódicos.

El modelado de la distribución de probabilidad de los parámetros se realiza a partir de un modelo de suma de M funciones de densidad Gaussianas, $p_i(\vec{x})$, cada una parametrizada por el vector de medias $Dx1$, $\vec{\mu}_i$, y una matriz de covarianzas DxD, σ_i;

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(\vec{x})$$

donde $p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sigma_i|^{1/2}}$ x $exp[-\frac{1}{2}(\vec{x} - \mu_i)^T \sum_i^{-1}(\vec{x} - \vec{\mu}_i)]$.

Los pesos de la mezcla, w_i, satisfacen la limitación $\sum_{i=1}^{M} w_i = 1$. El modelo se define como $\lambda = \{w_i, \vec{\mu}_i, \sigma_i\}$, donde $i = 1, ..., M$.

Normalmente se suelen usar matrices de covarianza diagonales por varias razones. Los GMMs con $M > 1$ con matrices de covarianza diagonales modelan distribuciones de vectores de características con elementos correlados. También GMMs con matrices diagonales son computacionalmente más eficientes que matrices de covarianza completas, las cuales requieren de repetidas inversiones de matrices DxD.

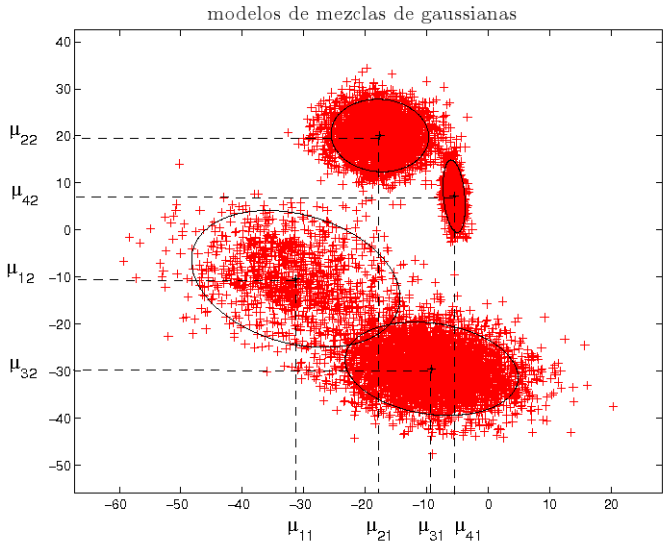

Figura 8: GMM bidimensional de 4 Gaussianas.

Dada una colección de vectores de entrenamiento, se estiman los parámetros de los modelos usando el algoritmo iterativo de máxima-expectación (EM, *Expectation-Maximization* en inglés) [16] (EM, expectation-maximization en inglés). Dicho algoritmo iterativamente refina los parámetros del GMM. Por ejemplo, para la iteración k y $k + 1$, $p(X|\lambda^{k+1}) > p(X|\lambda^k)$. Normalmente con 5 iteraciones es suficiente para la convergencia de los parámetros.

Para unos vectores de características desconocidos $\vec{X} = \{\vec{x}_1, ..., \vec{x}_T\}$ (se asumen que son independientes), el modelo GMM asigna una puntuación relacionada con su verosimilitud frente a un modelo λ que se calcula como:

$$log\ p(\vec{X}|\lambda) = \sum_{t=1}^{T} log\ p(\vec{x}_t|\lambda)$$

Existen dos hipótesis:

H_0: que el conjunto de vectores $\vec{X}$ pertenezca a la clase C.
H_1: que el conjunto de vectores $\vec{X}$ no pertenezca a la clase C.

Así, basándonos en el teorema de Bayes, la decisión óptima se toma a partir del cociente de las dos probabilidades:

$$\frac{p(\vec{X}|H_0)}{p(\vec{X}|H_1)}$$

Donde $p(\vec{X}|H_1)$ es la probabilidad de que la clase C no haya generado la muestra $\vec{x}$, y sin embargo haya sido cualquier otra clase.

Si dicho cociente supera un umbral θ, entonces se acepta la hipótesis H_0, sino se rechaza aceptando por lo tanto H_1.

Para estimar $p(\vec{x}|H_1)$ se hace uso de los que se conoce como modelos UBM (Universal Background Model). Un UBM es un modelo GMM estándar pero que ha sido entrenado a partir de observaciones de todos las clases (o un conjunto representativo de las mismas). Los UBM estiman la densidad de probabilidad de las observaciones, sobre todas las clases existentes. Por tanto, la verosimilitud frente al UBM mide la probabilidad de que la observación haya podido ser generada por una clase cualquiera.

En el sistema GMM UBM, el modelo se calcula mediante la adaptación de los parámetros de UBM usando los datos de entrenamiento de cada clase y un tipo de adaptación Bayesiana llamada estimación de máximo a posteriori (*MAP, maximum a posteriori*).

Los Modelos de mezclas Gaussianas son técnicas que originalmente fueron aplicadas al reconocimiento automático de locutor e idioma. El que dichas técnicas se hayan extendido al reconocimiento de emociones viene motivado por la similitud entre el reconocimiento de emoción e idioma y por los buenos resultados que los GMMs lograron en locutor e idioma. Así, se puede encontrar en la literatura gran cantidad de artículos que aplican el enfoque estadístico (generativo) en el reconocimiento de emociones en el habla [15] [17] [18].

3.7.2. SVM

Las Maquinas de Vectores Soporte (*SVM* o *Support Vector Machines*) son un tipo de clasificador de patrones binarios cuyo objetivo es asignar cada patrón a una clase [19]. A diferencia de los métodos tradicionales (generativos) los cuales modelan la probabilidad de una clase, los SVM son técnicas discriminativas, cuyo objetivo es modelar el plano de separación entre una clase y el conjunto de clases impostoras.

Planteamiento del problema de optimización

El problema consiste en construir un hiperplano de separación que divida el espacio R^n en dos regiones. Supongamos que tenemos dicho hiperplano, las muestras que caigan en una región pertenecerán a clase -1 y las que caigan en la otra a la clase 1. A este hiperplano se le conoce como hiperplano de separación.

Los vectores $\vec{x}$ que pertenecen al hiperplano de separación cumplirán la ecuación: $\vec{w} \cdot \vec{x}$+d=0, donde:

$\vec{w}$ es un vector normal al hiperplano de separación.
d es una constante.

La distancia $\frac{|d|}{||w||}$ es la distancia perpendicular desde el hiperplano al origen. Llamaremos d_+ y d_- a las distancias entre el hiperplano de separación y las muestras más cercanas a la clase +1 y -1 respectivamente. Con todo ello, el margen del hiperplano será la distancia entre las muestras más cercanas de las clases:

$$m = d_+ + d_-$$

Para el caso de datos linealmente separables, el objetivo es encontrar el hiperplano de separación que hace máximo este margen.

A la hora de formular formalmente el problema supondremos que todos los datos de entrenamiento cumplen una de las siguientes restricciones:

$$\vec{x}_i \cdot \vec{w} + d \geq +1 \text{ si } y_i = +1$$

$$\vec{x}_i \cdot \vec{w} + d \leq +1 \text{ si } y_i = -1$$

donde:

$y_i = \{1, -1\}$ representa la etiqueta de la clase a la que pertenece cada vector.
$i = \{1, ..., N\}$
N es el número de vectores de entrenamiento.

Combinando estas dos restricciones tenemos que:

$$\vec{y_i}(\vec{x}_i \cdot \vec{w}+\text{d})\text{-}1 \geq 0 \ \forall \text{i}$$

A los puntos más cercanos al hiperplano de separación se les conoce como vectores soporte, y están contenidos en los dos planos:

H1: $\vec{x}_i \cdot \vec{w}+\text{d}=+1$
H2: $\vec{x}_i \cdot \vec{w}+\text{d}=\text{-}1$

Ambos planos H1 y H2 son pararelos entre sí y a su vez paralelos al hiperplano de separación. Por lo tanto su componente normal seguirá siendo $\vec{w}$ [ver Figura 9] y sus respectivas distancias al origen serán:

$\frac{|1-b|}{||w||}$ para H1

$\frac{|-1-b|}{||w||}$ para H2

Cumpliendo todas las restricciones anteriores, las distancias d_+ y d_- serán $\frac{1}{||w||}$ por lo que el margen $\text{m}=\text{d}_+ + \text{d}_- = \frac{1}{||w||} + \frac{1}{||w||} = \frac{2}{||w||}$

El objetivo de los SVM es encontrar el hiperplano que maximice el margen de separación. Por lo tanto el problema se reduce a minimizar $||\vec{w}||$ sujeto a la restricción de:

$$\vec{y_i}(\vec{x}_i \cdot \vec{w} + d) - 1 \geq 0 \ \forall \text{i}.$$

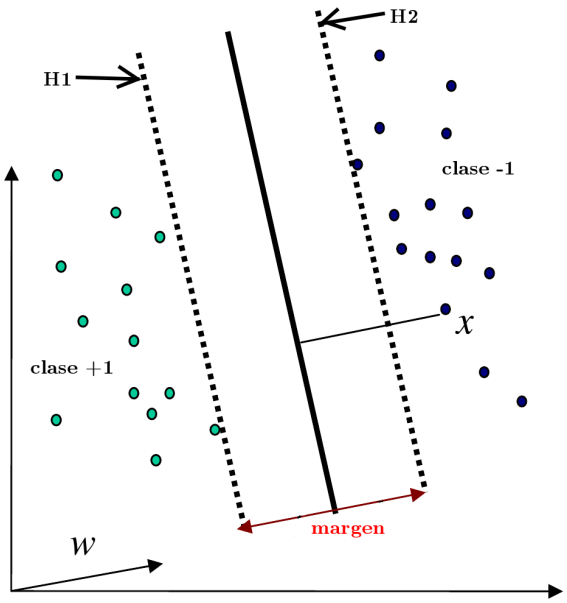

Figura 9: Concepto de un SVM.

Si los datos son linealmente separables, la resolución del problema obtiene un mínimo global, sino, el problema no es resoluble. Existen métodos computacionalmente eficientes para resolver problemas cuadráticos con múltiples restricciones lineales. Uno de ellos es mediante la formulación de Lagrange.

La formulación de Lagrange permite resolver un problema de optimización, como es nuestro caso, bajo una serie de restricciones mediante la introducción de unas nuevas variables, los multiplicadores de Lagrange, α_i. Puede demostrarse que es posible obtener el hiperplano óptimo de separación, $\vec{w}$, mediante una combinación lineal de los vectores soporte. El peso de cada uno de estos vectores se obtiene mediante los multiplicadores de Lagrange.

Como solución al problema se obtiene que el vector $\vec{w}$ se puede escribir en función de los vectores de entrenamiento, $\vec{x}_i$ como:

$$\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x}_i$$

Cada vector de entrenamiento tendrá asociado un multiplicador de Lagrange, α_i. Los vectores soporte tendrán un α_i asociado ≥ 0, mientras que el resto de vectores que no caen en los hiperplanos H1 o H2 tendrán un $\alpha_i = 0$ y por lo tanto no tendrán relevancia en el entrenamiento.

Clasificación del SVM

Una vez tenemos definido el hiperplano de separación entre las 2 clases, lo siguiente es encontrar una función que clasifique las muestras de test $\bar{\mathrm{x}}_t$ en su clase correspondiente. La función

$$f(\vec{x}_t) = \vec{w} \cdot \vec{x}_t + \mathrm{d} \Rightarrow f(\vec{x}_t) = \sum_{i=1}^{N} \alpha_i y_i \vec{x}_i \vec{x}_t + d$$

calcula la distancia del vector de test $\vec{x}_t$ al hiperplano de separación. Dicha función tomará valores positivos para las muestras pertenecientes a la clase +1 y negativos para las de la clase -1.

Se puede dar el caso en que algún vector ($\vec{x}_i$,y_i) viole la restricción

$$\vec{y}_i(\vec{x}_i \cdot \vec{w}+\mathrm{d})\text{-}1 \geq 0 \ \forall \mathrm{i}.$$

Para afrontar este problema lo que se debe hacer es relajar la restricción. Para ello se introduce unos márgenes de error h_i. $i = \{1, ..., N\}$ con $h_i \geq 0$ ∀i. La restricción será ahora $y_i(\vec{x}_i \cdot \vec{w} + d) \geq 1 - h_i$ con $i = \{1, ..., N\}$ con $h_i \geq 0$ ∀i.

Así, si $0 \leq h_i \leq 1$ la clasificación será correcta pero si a su vez $h_i > 0$, la muestra estará correctamente clasificada pero con un error asociado. Por otro lado, si $h_i \geq 1$, la clasificación será incorrecta.[Ver Figura 10]

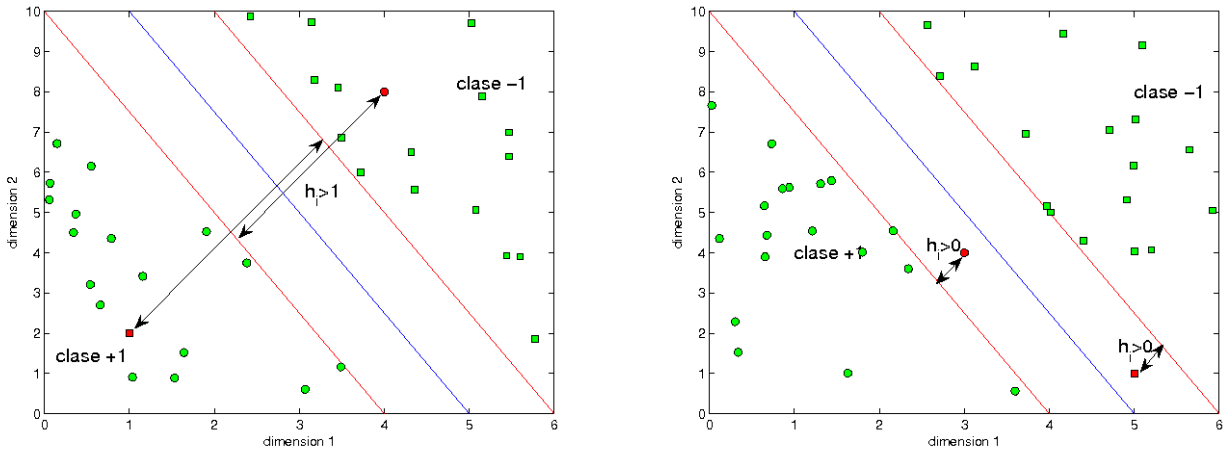

Figura 10: a) Muestras clasificadas incorrectamente con un valor h_i asociado. b) Muestras clasificadas correctamente pero con un error h_i.

Al añadir esta nueva variable pasaremos de uno a dos criterios a la hora de encontrar el hiperplano de separación:

• Maximizar el margen entre clases (criterio que ya teníamos anteriormente).
• Minimizar la función de pérdidas que será proporcional a las muestras incorrectamente clasificadas.

La relevancia de un criterio frente al otro se controla a través de una variable, a la que llamaremos coste, C. La variable coste será usada para dar más relevancia a un criterio frente al otro. Así, cuanto mayor sea el coste mayor importancia daremos a minimizar la función de pérdidas. Mientras que un valor pequeño de coste premiará en maximizar el margen entre clases. La variable coste será ajustada en la sección de pruebas para obtener los mejores resultados.

Hasta ahora hemos visto el funcionamiento de las Máquinas de Vectores Soporte en el modo de Clasificación (SVC) y para datos linealmente separables. Pero, ¿qué ocurre si los datos no cumplen esta premisa?

Separación no lineal de los datos

Un dato que hay que tener en cuenta es que, como se puede ver en la Figura 11, los datos que a priori no son separables en un espacio n-dimensional, sí pueden serlo en un espacio de mayor dimensión n'. Así por lo tanto, definiremos una función b($\vec{x}$) que mapea el espacio de

entrada n-dimensional (donde se sitúa $\vec{x}$) en un espacio de dimensión expandida n'

$$\mathrm{b}(\bar{\mathrm{x}}): R^n \rightarrow R^{n'}$$

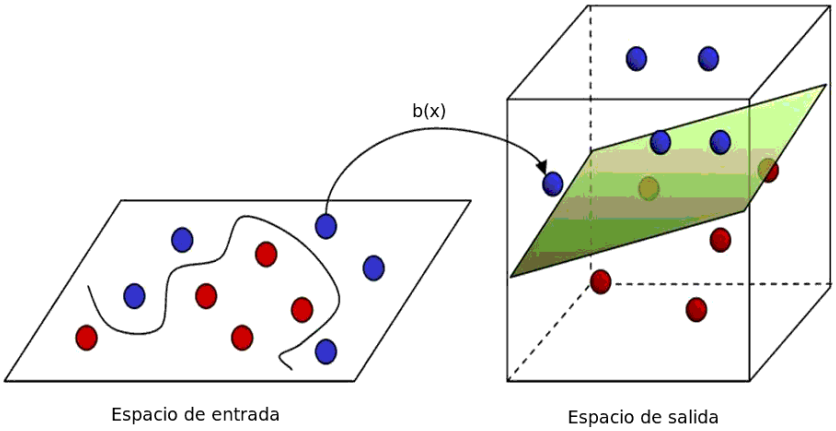

Figura 11: Mapeo de los vectores $\vec{x}$ 2-dimensionales a b($\vec{x}$) 3-dimensionales.

Este es el momento de introducir la función kernel. Esta función nos permite calcular el producto interno de dos vectores sin necesidad de conocer explícitamente el vector mapeo en el espacio transformado.

$$\mathrm{K}(\vec{x}_i, \vec{x}_j) = \mathrm{b}(\vec{x}_i) \cdot \mathrm{b}(\vec{x}_j)$$

A la hora de elegir la función kernel, ésta debe de satisfacer el teorema de Mercer. El teorema de Mercer nos dice si un kernel $K(\cdot,\cdot)$ cumple las propiedades del producto escalar y por lo tanto útil para un SVM. No nos dice sin embargo como construir dicha función $K(\cdot,\cdot)$.

La elección de una buena función kernel debe satisfacer dos premisas. Debe ser tal, que dadas dos locuciones $\vec{x}_i$ y $\vec{x}_j$, obtenga un valor de similitud entre ambas. También debe de ser computacionalmente eficiente ya que durante el proceso de entrenamiento y test se van a llevar a cabo muchos productos internos.

Las Máquinas de Vectores Soporte es una herramienta novedosa que ha aparecido en la última década en la clasificación automática de patrones. Ha llegado a ser muy popular debido a su capacidad de solventar muchos de los problemas de los ANNs (*Artificial Neural Networks*) y de los HMMs (*Hidden Markov Models*) gracias a su efectiva capacidad de discriminación. En contraposición con los ANNs, tienen la ventaja de tratar con muestras de muy alta dimensión. Estas características han hecho a los SVMs muy populares y existosos en muchos campos de aplicación. No obstante, existen algunas limitaciones a la hora de usar los SVMs. Una de estas limitaciones es que los SVMs están restringidos a trabajar con vectores de entrada de longitud fija. Otra limitación es que los SVMs sólo clasifican, pero no dan una medida fiable de la probabilidad de la correcta o incorrecta clasificación.

Los SVMs presentan muy buen rendimiento en tareas de procesado vocal como reconocimiento de idioma y locutor. Es por eso por lo que también se usan para reconocimiento automático de emociones en el habla y como muestra de ello se pueden ver [20], [17], [21] y [22] donde se usan los rasgos acústicos y prosódicos del habla para modelar los SVMs.

3.7.3. SVMs basados en supervectores GMMs

Los SVMs basados en supervectores GMMs son técnicas de clasificación de patrones que aunan las ventajas de los sistemas generativos, como son los GMMs, con las de los sistemas discriminativos como son los SVMs [23].

Un supervector GMM se construye apilando los vectores medios d-dimensionales de las M componentes gaussianas. El supervector GMM puede ser considerado como una función kernel $SV(\vec{x})$ que mapea los vectores de características $\vec{x}$ en un vector de mayor dimensión $L = M * d$. En este espacio L-dimensional del supervector es donde se entrena un SVM para así conseguir un modelo $\vec{w_e}$. Para este caso, la función de puntuación $s'(\vec{w_e}, SV(\vec{x_{test}}))$ se define como:

$$s'(\vec{w_e}, SV(\vec{x_{test}})) = \vec{w_e} * SV(\vec{x_{test}})^T$$

Suponemos que tenemos un modelo de UBM el cual es adaptado (MAP) a partir de los vectores de parámetros de una locución. Dicha adaptación conforma un modelo de mezclas gaussianas definido como:

$\lambda = \{w_i, \mu_i, \sigma_i\}$, donde $i = 1, ..., M$ con M el número de mezclas unimodales Gaussianas. A partir de este modelo, se forma el supervector GMM. Este proceso se muestra en la Figura 12

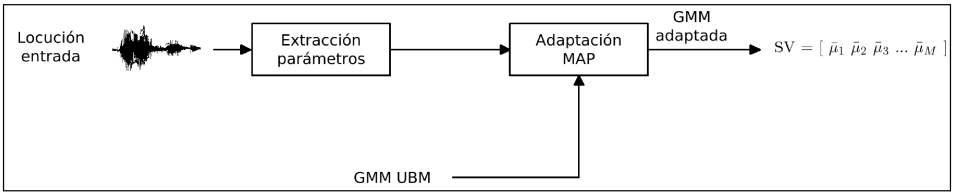

Figura 12: Construcción de un supervector GMM a partir de una locución de voz.

Como ejemplo de construcción de un supervector GMM podemos ver la Figura 13 donde $d = 2$, $M = 3$ y $L = M * d = 6$. En este caso, vectores de parámetros bidimensionales modelan 3 componentes gaussianas. Como se puede ver, los vectores medios bidimensionales de las 3 componentes gaussianas conforman el supervector $SV = [\vec{\mu_1}\ \vec{\mu_2}\ \vec{\mu_3}] = [\mu_{11}\ \mu_{12}\ \mu_{21}\ \mu_{22}\ \mu_{31}\ \mu_{32}]$

Se ha visto que esta técnica de SVM basados en supervectores GMM ha dado excelentes resultados en tareas de reconocimiento de locutor [24] e idioma usando el nivel acústico del habla. A parte del reconocimiento de locutor e idioma, también se ha extendido al reconocimiento de emociones. Así, [23] propone un SVM basado en supervectores GMMs a partir de rasgos espectrales mientras que en [25] lo proponemos a partir de rasgos prosódicos del habla para el reconocimiento de emociones.

3.7.4. *Anchor Models*

El espacio de proyección de los *Anchor Models* es una función que mapea cada locución de habla desde el espacio de características original en un nuevo espacio *anchor model*. Las dimensiones de este nuevo espacio son puntuaciones de similitud de cada locución frente a modelos previamente entrenados $\vec{m} = \{m_1...m_N\}$. Estos modelos han sido entrenado mediante técnicas de clasificación como GMMs, SVMs, etc. Este espacio de similitud permite obtener el comportamiento de una locución $\vec{x}$ frente a los modelos $\vec{m}$ obteniendo así un vector de puntuaciones de similitud:

$$\vec{S_x} = [s_{x,m_1}...s_{x,m_N}]$$

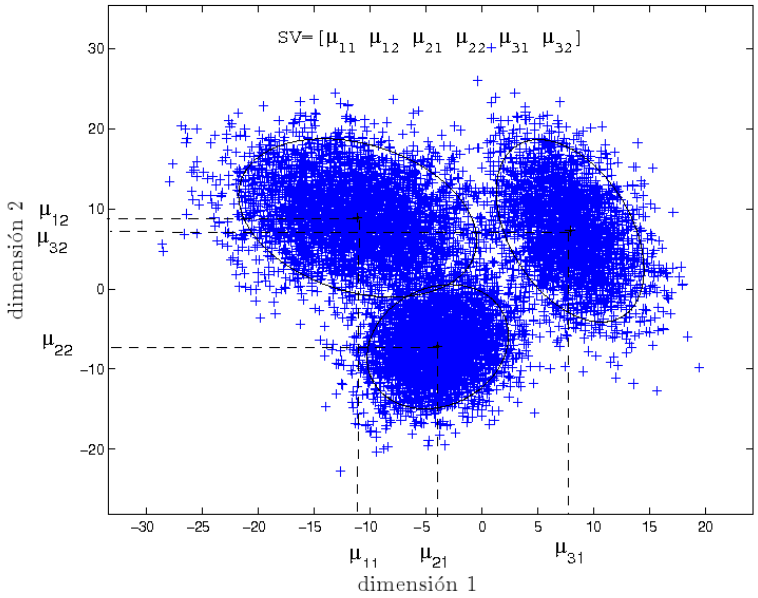

Figura 13: Ejemplo de construcción de un supervector GMM a partir de 3 mezclas gaussianas bidimensionales.

donde se apilan las puntuaciones individuales del vector $\vec{x}$ frente a cada uno de los modelos m_i [Figura 14].

A partir de entonces, se puede considerar el vector $\vec{S}_{x,m}$ como el vector de parámetros de la locución $\vec{x}$ y un nuevo modelo $m_i^{'}$ puede ser generado en el espacio del *anchor model* usando técnicas de aprendizaje como GMMs, SVMs, n-gramas, etc.

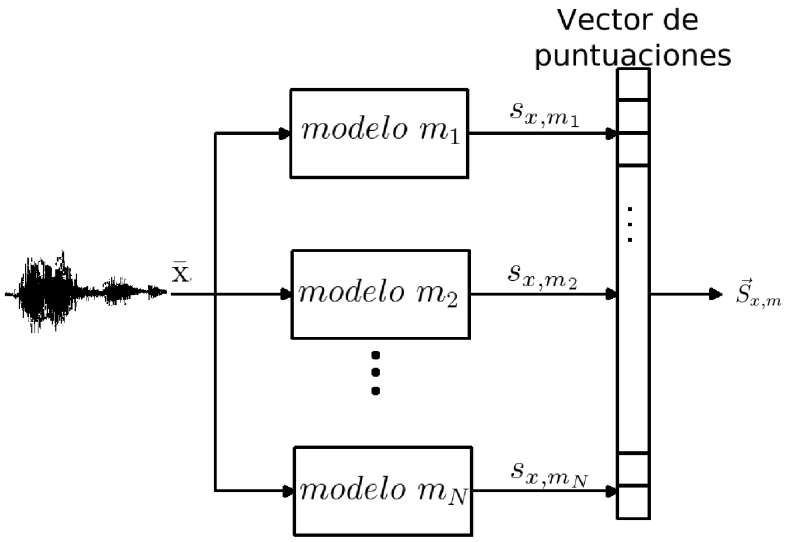

Figura 14: $\vec{S}_x$ agrupa las puntuaciones de similitud del vector $\vec{x}$ frente a cada modelo m_i.

El valor de N define la dimensión del espacio de los anchor models y la puntuación del vector $\vec{x}$ frente a cada modelo $\vec{m}_i$ define la distancia a cada uno de los ejes de este nuevo espacio dimensional. De la teoría de Vapnik-Chervonenkis [26] se deduce que cuanto mayor sea el valor de N, mayor dimensión será el espacio de características del anchor model y por ello más fácil será encontrar un comportamiento característico de la locución $\vec{x}$. En el reconocimiento de emociones N estará limitada por el número de emociones disponibles.

Fusión de *Anchor Models*

La función de similitud o puntuación s_{x,m_i} nos ofrece una medida de similitud entre el vector $\vec{x}$ y el modelo $\vec{m}_i$. Cada técnica usada para construir los modelos usa una función de similitud diferente. Así, por ejemplo, los SVMs usan la distancia algebraica mientras que los GMMs, como ya hemos visto, usan un criterio de similitud estadística $\frac{p(\vec{X}|H_0)}{p(\vec{X}|H_1)}$. Mediante el uso de varias funciones de similitud s_{x,m_i}, la información obtenida puede ser complementaria y con ello se puede obtener una mejora de los resultados.

La fusión de anchor models (en inglés AMF, *Anchor Model Fusion*) es una técnica novedosa ideada por el ATVS [27], [28] que ha logrado dar muy buenos resultados pues obtiene información complementaria procedente de varios subsistemas. Consiste en usar varias técnicas de entrenamiento (y con ello varias funciones de similitud) como pueden ser los SVMs, GMMs, etc. para generar los modelos $\vec{m}_i$.

En el caso de reconocimiento de emociones, el vector $\vec{m}$ incluyen los n modelos de emociones pre-entrenadas por cada uno de los sistemas de reconocimiento de emociones a fusionar. Así, el vector de parámetros generado a partir de las puntuaciones de la locución $\vec{x}$ frente a cada modelo de $\vec{m}$ por cada uno de los N_{sist} sistemas es:

$$\vec{S}_{m,x} = [\vec{S}^1_{m,x}, ..., \vec{S}^{N_{sist}}_{m,x}]$$

La Figura 15 muestra una versión esquemática de AMF. Para este caso, la dimensión del espacio de los anchor models es $N = n * N_{sist}$.

3.7.5. Otras: LDA, HMM

Análisis de Discriminación Lineal

El Análisis de Discriminación Lineal (en inglés LDA, *Linear discriminant analysis*) y la discriminación lineal de Fisher relacionada son métodos usados en estadística y en aprendizaje automático cuyo objetivo es encontrar la combinación lineal de características que mejor separa 2 o más clases.

LDA está muy relacionado con ANOVA (análisis de varianza) y con el análisis en regresión, que también intentan expresar una variable como combinación lineal de otros rasgos o características. Mientras que en estos dos últimos métodos la variable dependiente se cuantifica numéricamente, en LDA es una variables categórica (por ejemplo, la clase *emoción 1*).

LDA ha sido usado satisfactoriamente como técnica de reducción dimensional en muchos problemas de clasificación, como reconocimiento de habla, reconocimiento de cara o recuperación de información multimedia. En [20] se usa LDA como clasificador de emociones.

El reconocimiento de emociones en el habla es un pequeño ejemplo de las aplicaciones en las que se hace uso de la técnica LDA. Otra es el reconocimiento facial. Cada cara se representa por un gran número de valores de píxeles. En este caso se usa LDA para reducir el número de características a un número más manejable antes de la clasificación. Cada dimensión nueva es combinación lineal de los valores de los píxeles. [29]

Éstos son sólo dos ejemplos de las innumerables tareas en las que la aplicación de LDA puede emplearse con éxito.

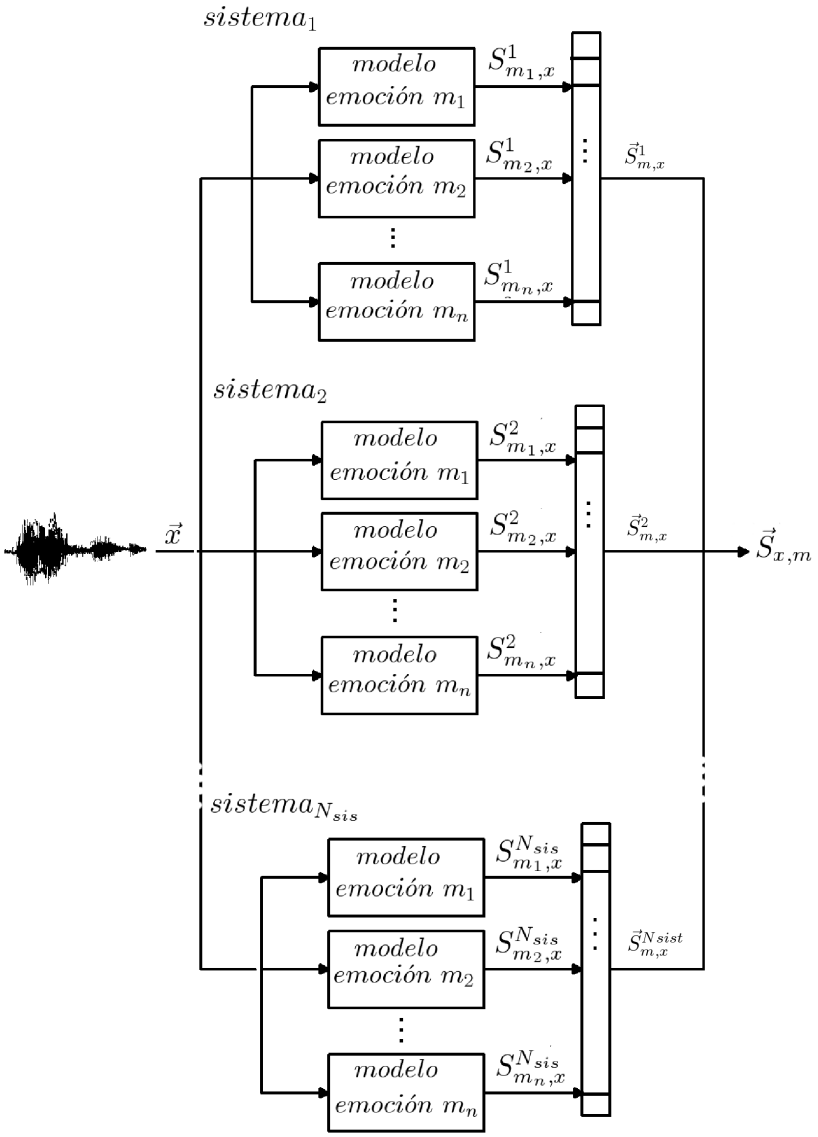

Figura 15: Diagrama de funcionamiento del AMF. El vector de parámetros final de la locución $\vec{x}$ es la concatenación de las puntuaciones de similitud de $\vec{x}$ frente a cada modelo de emoción m_i para cada uno de los N_{sist} sistemas.

Modelos Ocultos de Markov

Un HMM (en inglés, *Hidden Markov Models*) o modelo oculto de Markov es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u ocultos, de ahí el nombre) de dicha cadena a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo sucesivos análisis, por ejemplo en aplicaciones de reconocimiento de patrones. Un HMM se puede considerar como la red bayesiana dinámica más simple.

En un modelo de Markov normal, el estado es visible directamente para el observador, por lo que las probabilidades de transición entre estados son los únicos parámetros. En un modelo oculto de Markov, el estado no es visible directamente, sino que sólo lo son las variables influidas por el estado. Cada estado tiene una distribución de probabilidad sobre los posibles símbolos de salida. Consecuentemente, la secuencia de símbolos generada por un HMM proporciona cierta información acerca de la secuencia de estados.

Los modelos ocultos de Markov son especialmente aplicados a reconocimiento de formas

temporales, como reconocimiento del habla, de escritura manual, de gestos, etiquetado gramatical o en bioinformática.

Dado el buen funcionamiento de esta técnica en tareas como el reconocimiento de habla, también se ha aplicado al reconocimiento de emociones [20] combinado con otras técnicas de clasificación como los GMM o SVM [30].

4

Diseño y Desarrollo

Este capítulo comienza haciendo un análisis de las bases de datos de voz emocional existentes en la literatura. Además se describen las que han sido utilizadas en este trabajo: *SUSAS Simulated*, *SUSAS Actual* y *Ahumada III*.

También se detallan los procedimientos seguido para la creación y evaluación de cada uno de los sistemas propuestos: parametrización del audio, entrenamiento de los modelos y su posterior evaluación.

4.1. Medios disponibles (BBDD, software, máquinas...)

4.1.1. Bases de Datos Utilizadas

Para poder evaluar nuestros sistemas de reconocimiento de emociones en el habla se necesitan bases de datos sobre las que testearlos. Cuanto mayor sea la diversidad de la base de datos, más realísticos serán los resultados obtenidos.

En la Tabla 2 aparece una colección de las principales bases de datos de habla emocional existentes para tareas de reconocimiento y síntesis de emociones. En ella se describe información sobre cada base de datos como: idioma, locutores, emociones existentes, etc.

Referencia	**Idioma**	**Sujetos**	**Otras señales**	**Propósito**	**Emociones**	**Tipo de datos**
Abelin and Allwood (2000)	Sueco	1 Nativo	–	Reconoc.	Eo, Mo, Ag, Tz, Se, Dt, Dom, Tz	Simulados
Alpert et al. (2001)	Inglés	22 Pacientes 19 sanos	–	Reconoc.	Dn, Nl	Natural
Alter et al. (2000)	Alemán	1 Female	EEG	Reconoc.	Eo, Fd, Nl	Simulados
Ambrus (2000) Interface	Inglés, Eslovaco	8 Actores	LG	Síntesis	Eo, Dt, Mo, Nl, Se	Simulados
Amir et al. (2000)	Hebreo	40 Estudiantes	LG,M,G,H	Reconoc.	Eo, Dt, Mo, Ag, Tz	Natural
Ang et al. (2002)	Inglés	Muchos	–	Reconoc.	An, Dn, Nl, Fd, Co	Natural
Banse and Scherer (1996)	Alemán	12 Actores	V	Reconoc.	C/F Eo, Fd, Tz,...	Simulados
Batliner et al. (2004)	Alemán, Inglés	51 Niños	–	Reconoc.	Eo, Ao, Ag, Se	Provocados
Bulut et al. (2002)	Inglés	1 Actress	–	Síntesis	Eo, Fd, Nl, Tz	Simulados
Burkhardt and Sendlmeier (2000)	Alemán	10 Actores	V, LG	Síntesis	Eo, Mo, Ag, Nl, Tz, Ao, Dt	Simulados
Caldognetto et al. (2004)	Italiano	1 Nativo	V, IR	Síntesis	Eo, Dt, Mo, Ag, Tz, Se	Simulados
Choukri (2003), Groningen	Holandés	238 Nativos	LG	Reconoc.	Desconocidas	Simulados
Chuang and Wu (2002)	Chino	2 Actores	–	Reconoc.	Eo, Aa, Fd, Mo, Se, Tz	Simulados
Clavel et al. (2004)	Inglés	18 de la TV	–	Reconoc.	Nl, niveles de Mo	Simulados

Tabla 2 – continúa de la página anterior

Referencia	Idioma	Sujetos	Otras señales	Propósito	Emociones	Tipo de datos
Cole (2005), Kids' Speech	Inglés	780 Niños	V	Reconoc., Síntesis	Desconocidas	Natural
Cowie and Douglas-Cowie (1996), Belfast Structured	Inglés	40 Nativos	–	Reconoc.	Eo, Mo, Fd, Nl, Tz	Natural
Douglas-Cowie et al. (2003), Belfast Natural	Inglés	125 de la TV	V	Reconoc.	Varias	Semi-natural
Edgington (1997)	Inglés	1 Actor	LG	Síntesis	Eo, Ao, Mo, Fd, Nl, Tz	Simulados
Engberg and Hansen (1996), DES	Danish	4 Actores	–	Síntesis	Eo, Fd, Nl, Tz, Se	Simulados
Fernandez and Picard (2003)	Inglés	4 Drivers	–	Reconoc.	Nl, Ss	Natural
Fischer (1999), Verbmobil	Alemán	58 Nativos	–	Reconoc.	Eo, Dn, Nl	Natural
France et al. (2000)	Inglés	70 Pacientes, 40 sanos	–	Reconoc.	Dn, Nl	Natural
Gonzalez (1999)	Inglés, Español	Desconocidos	–	Reconoc.	Dn, Nl	Provocados
Hansen (1996), SUSAS	Inglés	32 Varios	–	Reconoc.	Eo, Ld eff., Ss, Tl	Natural, simulated
Hansen (1996), SUSC-0	Inglés	18 No nativos	H,PS,R	Reconoc.	Nl, Ss	A-estrés
Hansen (1996), SUSC-1	Inglés	20 Nativos	–	Reconoc.	Nl, Ss	P-estrés
Hansen (1996), DLP	Inglés	15 Nativos	–	Reconoc.	Nl, Ss	C-estrés
Hansen (1996), DCIEM	Inglés	Desconocidos	–	Reconoc.	Nl, privación de sueño	Provocados

Tabla 2 – continúa de la página anterior

Referencia	Idioma	Sujetos	Otras señales	Propósito	Emociones	Tipo de datos
Heuft et al. (1996)	Alemán	3 Nativos	–	Síntesis	Eo, Mo, Ag, Tz,...	Simulados, provocados
Iida et al. (2000), ESC	Japonés	2 Nativos	–	Síntesis	Eo, Ag, Tz	Simulados
Iriondo et al. (2000)	Español	8 Actores	–	Síntesis	Mo, Ag, Tz, Se,...	Simulados
Kawanami et al. (2003)	Japonés	2 Actores	–	Síntesis	Eo, Fd, Nl, Tz	Simulados
Lee and Narayanan (2005)	Inglés	Desconocidos	–	Reconoc.	Negat.–Posit.	Natural
Liberman (2005), Emotional Prosody	Inglés	Actores	–	Desconocido	Ad, C/F Eo, Fd, Nl, Pc, Tz, Se,...	Simulados
Linnankoski et al. (2005)	Inglés	13 Nativos	–	Reconoc.	An, Eo, Mo, Tz,...	Provocados
Lloyd (1999)	Inglés	1 Nativo	–	Reconoc.	Stress fonológico	Simulados
Makarova and Petrushin (2002), RUSSLANA	Ruso	61 Nativos	–	Reconoc.	Eo, Fd, Se, Tz, Mo, Nl	Simulados
Martins et al. (1998), BDFALA	Portugués	10 Nativos	–	Reconoc.	Eo, Dt, Fd, Iy	Simulados
McMahon et al. (2003), ORESTEIA	Inglés	29 Nativos	–	Reconoc.	Ma, Sk, Ss	Provocados
Montanari et al. (2004)	Inglés	15 Niños	V	Reconoc.	Desconocidas	Natural
Montero et al. (1999), SES	Español	1 Actor	–	Síntesis	Eo, Dt, Fd, Tz	Simulados
Mozziconacci and Hermes (1997)	Holandés	3 Nativos	–	Reconoc.	Eo, Ao, Mo, Ag, Iy, Nl, Tz	Simulados
Niimi et al. (2001)	Japonés	1 Male	–	Síntesis	Eo, Ag, Tz	Simulados

Tabla 2 – continúa de la página anterior

Referencia	Idioma	Sujetos	Otras señales	Propósito	Emociones	Tipo de datos
Nordstrand et al. (2004)	Sueco	1 Nativo	V, IR	Síntesis	Fd, Nl	Simulados
Nwe et al. (2003)	Chino	12 Nativos	–	Reconoc.	Eo, Mo, Dt, Ag,...	Simulados
Pereira (2000)	Inglés	2 Actores	–	Reconoc.	C/F Eo, Fd, Nl, Tz	Simulados
Petrushin (1999)	Inglés	30 Nativos	–	Reconoc.	Eo, Mo, Fd, Nl, Tz	Simulados, Natural
Polzin and Waibel (2000)	Inglés	Desconocidos	–	Reconoc.	Eo, Mo, Nl, Tz	Simulados
Polzin and Waibel (1998)	Inglés	5 estudiantes de drama	LG	Reconoc.	Eo, Mo, Fd, Nl, Tz	Simulados
Rahurkar and Hansen (2002), SOQ	Inglés	6 soldados	H, R, PS, ES	Reconoc.	5 niveles de estreś	Natural
Scherer (2000b), Lost Luggage	Varios	109 Passengers	V	Reconoc.	Eo, Hr, Ie, Tz, Ss	Natural
Scherer (2000a)	Alemán	4 Actores	–	Ecological	Eo, Dt, Mo, Ag, Tz	Simulados
Scherer et al. (2002)	Inglés, Alemán	100 Nativos	–	Reconoc.	2 Tl, 2 Ss	Natural
Schiel et al. (2002), SmartKom	Alemán	45 Nativos	V	Reconoc.	Eo, In, Nl	Natural
Schroder and Grice (2003)	Alemán	1 Male	–	Síntesis	Soft, modal, loud	Simulados
Schroder (2000)	Alemán	6 Nativos	–	Reconoc.	Eo, Ao, Dt, Pn,...	Simulados
Slaney and McRoberts (2003), Babyears	Inglés	12 Nativos	–	Reconoc.	An, An, Pn	Natural
Stibbard (2000), Leeds	Inglés	Desconocidos	–	Reconoc.	Amplio rango	Natural, elicited
Tato (2002), AIBO	Alemán	14 Nativos	–	Síntesis	Eo, Ao, Fd, Nl, Tz	Provocados
Tolkmitt and Scherer (1986)	Alemán	60 Nativos	–	Reconoc.	Cognitive Ss	Provocados

Tabla 2 – continúa de la página anterior

Referencia	Idioma	Sujetos	Otras señales	Propósito	Emociones	Tipo de datos
Wendt and Scheich (2002), Magdeburger	Alemán	2 Actores	–	Reconoc.	Eo, Dt, Mo, Fd, Tz	Simulados
Yildirim et al. (2004)	Inglés	1 Actriz	–	Reconoc.	Eo, Fd, Nl, Tz	Simulados
Yu et al. (2001)	Chino	Nativos	–	Reconoc.	Eo, Fd, Nl, Tz	Simulados
Yuan (2002)	Chino	9 Nativos	–	Reconoc.	Eo, Mo, Ag, Nl, Tz	Provocados

Tabla 2: Recopilación de bases de datos de habla emocional. Tabla adaptada de [2]. Abreviaturas de emociones: Dn: Diversión, Aa: Antipatía, Eo: Enfado, Ma: Molestia, An: Aprobación, An: Atención, Ad: Ansiedad, Ao: Aburrimiento, In: Insatisfación, Dom: Dominio, Dn: Depresión, Dt: Disgusto, Fd: Frustración, Mo: Miedo, Fd: Felicidad, Ie: Indiferencia, Iy: Ironía, Ag: Alegría, Nl: Neutra, Pc: Pánico, Pn: Prohibición, Se: Sorpresa, Tz: Tristeza, Ss: Estrés, Tz: Timidez, Sk: Shock, Co: Cansancio, Tl: Tarea con carga de estrés, Pn: Preocupación. Abreviaturas para otras señales: PS: Presión sanguínea, ES: Examinación de sangrue, EEG: Electroencefalograma, G: Respuesta cutánea galvánica, H: Tasa latido corazón, IR: Cámara infrarroja, LG: Laringógrafo, M: Miograma de la cara, R: Respiración, V: Video. Otras abreviaturas: C/F: Caliente/Frio, Ld eff.: efecto Lombard, A-stress, P-stress, C-stress: stress Real, Físico y Cognitivo, respectivamente, Sim.: Simulado, Prov.:Provocado, N/A: No disponible.

Para el entrenamiento de los modelos y su posterior evaluación haremos uso de 2 bases de datos disponibles en el ATVS (contacto: atvs@uam.es) como son *SUSAS* (en inglés, *Speech Under Simulated and Actual Stress*) y Ahumada III.

SUSAS: Speech Under Simulated and Actual Stress

Speech Under Simulated and Actual Stress (SUSAS) [31] es una base de datos en inglés que ha sido empleada con frecuencia en el estudio de la síntesis y reconocimiento de habla bajo condiciones de estrés [20]. Esta base de datos fue originalmente diseñada por John H.L. Hansen en 1998 para tareas de reconocimiento de habla bajo estrés. En el grupo ATVS esta base de datos ha sido obtenida del LDC (*Linguistic Data Consortium*) [32].

Se ha elegido la base de datos SUSAS por las siguientes razones:

- contiene un gran número de emociones.
- permite hacer comparaciones con anteriores trabajos.
- se dispone de los IDs de los locutores.
- existen datos de tanto habla real como simulada.

Buena parte de la literatura existente sobre el reconocimiento de emociones en el habla usa la base de datos SUSAS para llevar a cabo sus experimentos [17] [10] [30]. Todos los ficheros de voz de SUSAS están muestreados a 8KHz y con 16 bits por muestra. La base de datos consta de dos tipos de datos según éstos sean simulados o reales. Así, tenemos una parte llamada *SUSAS Simulated* y otro llamada *SUSAS Actual* respectivamente.

SUSAS Simulated contiene habla simulada de 9 locutores (todos hombres) y 11 estilos de habla. Los 9 locutores se distribuyen en 3 grupos con (i) acento general de USA (*g1*, *g2*, *g3*), (ii) acento de Nueva Inglaterra/Boston (*b1*, *b2*, *b3*), y (iii) acento de la ciudad de Nueva York (*n1*, *n2*, *n3*). Los datos incluyen 8 estilos: *angry (a), clear (c), fast (f), loud (l), neutral (n), question (q), slow (s), soft (w)* y otros 3 estilos bajo diferente grado de estrés: *lombard (lom), cond70 (c70), cond50 (c50). angry* corresponde a un estilo de habla enfadado, *clear* a habla con una clara pronunciación, *fast* a habla rápida, *loud* es habla enérgica, *neutral* es un estilo de habla normal o neutra, *question* corresponde a habla con entonación interrogativa, *slow* es habla lenta y *soft* habla suave o poco enérgica. El estilo de habla *lombard* se produce como consecuencia del efecto *Lombard* que consiste en la tendencia involuntaria de los locutores en elevar la intensidad de voz cuando se encuentran en un ambiente altamente ruidoso para mejorar su audibilidad. Las condiciones de estrés *cond50* y *cond70* corresponden a habla producida por locutores mientras realizan una tarea estresante con un *joy-stic* en un ordenador. Según el grado de dificultad, bajo o alto, tendremos los estilos de habla *cond50* y *cond70* respectivamente.

SUSAS Actual contiene habla real de 7 locutores (3 mujeres y 4 hombres) y 5 condiciones de estrés: *neutral (n), medst (m), hist (h), freefall (f), scream (s)*. Los 4 locutores masculinos se denotan como *m1*, *m2*, *m3* y *m4*, mientras que los 3 femeninos como *f1*, *f2* y *f3*. La condiciones de haba bajo estrés *medst* y *hist* corresponden a habla en condiciones en que los locutores están realizando una tarea que les supone un estrés. Dependiendo de si el grado de estrés es moderado (*moderate*) o alto (*high*) tendremos los estilos *medst* y *hist* respectivamente. Por otra parte, los estilos de habla *freefall, scream y neutral* se obtienen de locutores montados en atracciones de un parque temático. *freefall* se consigue recogiendo voz mientras los locutores se montan en una montaña rusa y *scream* mientras lo hacen en una atracción de miedo.

Los datos de *Simulated* y *Actual* consisten en locuciones de palabras pertenecientes a un conjunto de 35 palabras (*break, change, ...*). Cada palabra dispone de 2 realizaciones por locutor y emoción.

Un ejemplo de la primera de las dos repeticiones de una locución de la base de datos *SUSAS Simulated* de la palabra *break* bajo el estilo de habla *angry* del locutor *b2* lo tenemos en la Figura 16.

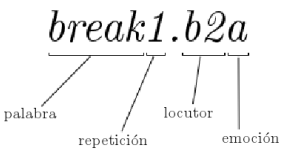

Figura 16: Ejemplo de una locución de la base de datos *SUSAS Simulated*.

Ahumada III (*Ah3R1*)

Ahumada III es una base de datos de habla en español descrita en [33] recogida de casos forenses reales por el Departamento de Procesado de Audio e Imagen de la Guardia Civil Española. Su versión actual, *Ahumada III Release 1* (*Ah3R1*) incluye habla de casos forenses obtenidos usando el sistema típico de grabación de la Guardia Civil, cintas analógicas magnéticas con grabaciones GSM. Tambíen usando SITEL, un sistema español de interceptación legal

de las telecomunicaciones. Este sistema graba conversaciones telefónicas digitales conectado directamente a todos los operadores telefónicos.

Ah3R1 incluye gran variabilidad de condiciones, como ruido, características del entorno, estado anímico, país, región de origen y dialecto de los locutores, etc.

En la última década, la Guardia Civil ha ido creando una serie de base de datos con el propósito de hacer sistemas más robustos mediante la ampliación de la variabilidad de las condiciones. Como ejemplos de dichas bases de datos tenemos: Ahumada I [34], Gaudi (2001), Baeza (2004-2006) o Ahumada II.

El tamaño esperado de *Ah3R1* es muy grande tanto en el número de llamadas disponibles como en el número de locutores. Sin embargo, como las condiciones no son uniformes y las grabaciones de voz tienen que estar autorizadas una por una, se espera que progresivamente vayan estando disponibles diferentes versiones de la base de datos.

Ah3R1 contiene datos de 69 locutores sacados de casos reales en llamadas GSM BDRA en España con variedad en el país de origen de los locutores, del estado emocional, condiciones acústicas y dialectos. En el único caso en que no hay variabilidad es en el género, pues los 69 locutores son hombres. Para cada locutor existen dos minutos de habla disponibles, los cuales se usan para el entrenamiento de los modelos que caractericen el habla de dicho locutor. Además, para tareas de evaluación se dispone de 10 segmentos de habla para los 31 primeros locutores y cinco para los 38 restantes, cada uno de diferentes llamadas telefónicas. Dichos fragmentos constan de entre 7 y 25 segundos de habla, con una duración media de 13 segundos.

Los estilos de habla contenidos en *Ah3R1* son *neutro-bajo*, *neutro*, *neutro-exaltado* y *exaltado*. En la Figura 17 vemos un ejemplo de un par de locuciones de *Ah3R1* perteneciente al locutor 23.

Un ejemplo de dos locuciones de *Ah3R1*, una de entrenamiento y otra de test se puede ver en la Figura 17.

Figura 17: a) Locución de *Ah3R1* de entrenamiento (120sg) del locutor 23 y emoción *neutro-exaltado*. b) Locución número 4 de test de *Ah3R1* del locutor 23 y emoción *neutro*.

Los datos de *Ah3R1* son públicos y su acceso está disponible para proyectos de investigación mediante una licencia que debe ser firmada por la Guardia Civil. (contacto: crim-acustica@guardiacivil.es). Varias muestras de segmentos de habla se puede escuchar directamente en la página web del ATVS (http://atvs.ii.uam.es/) para así percibir la calidad y variedad de las grabaciones de *Ah3R1*.

4.1.2. Software y Máquinas

El hardware utilizado para el desarrollo de este proyecto ha sido un ordenador de uso personal con procesador Intel Pentium IV y SO Debian y distribución Ubuntu. También he tenido acceso a los ordenadores del resto de grupo de trabajo y al rack de servidores para lanzar las pruebas.

Todos estos medios fueron suministrados por el grupo ATVS de la Universidad Autónoma de Madrid (UAM).

4.2. Diseño

4.2.1. Parametrización del audio

El primer paso a la hora de implementar un sistema de reconocimiento de habla es la extracción de los rasgos característicos de la señal de voz que la identifiquen frente al resto. A esto proceso se le llama parametrización y su variedad es muy extensa dependiendo de la tarea que se pretenda realizar. Así, según el nivel de la voz en que trabajen, tenemos la parametrización acústica y la parametrización prosódica que son dos de las mas importante y utilizadas.

La extracción de rasgos de bajo nivel como son los rasgos acústicos se utiliza normalmente para modelar el comportamiento del locutor. Este tipo de rasgos se suele utilizar para autentificación de locutor porque los locutores tienen menos control sobre los detalles espectrales del habla que sobre rasgos de alto nivel como el pitch. Como ejemplo de parametrización acústica están los MFCC (*Mel Frequency Cepstral Coefficients*), SDC (*Shifted Delta Cepstral*) o LFPC (*Low frecuency power coefficients*).

La prosodia es una rama de la lingüística que analiza y representa formalmente aquellos elementos de la expresión oral, tales como el acento, los tonos y la entonación. Su manifestación concreta en la producción de la palabra se asocia de este modo a las variaciones de la frecuencia fundamental, de la duración y de la intensidad que constituyen los parámetros prosódicos físicos.

Parametrización prosódica

En la literatura existen muchos trabajos que han encontrado relación entre las variaciones de la prosodia del locutor y la información de su estado emocional [10], [35]. Muchos sistemas de reconocimiento de emociones utilizan los rasgos prosódicos del habla como señal de entrada. Los rasgos prosódicos más comúnmente utilizados son la frecuencia fundamental o *pitch* (F0), la energía y sus correspondientes velocidades, también conocidas como rasgos Δ y la duración.

Se va a hacer uso de la parametrización prosódica para la realización de nuestros sistemas. En concreto, la señal de audio es enventanada cada 10 ms usando una ventana de Hamming de 40 ms [Figura 18]. Mediante la herramienta *Praat* [36] se extrae por cada ventana la *energía* y el *log F0* obteniendo un vector de energías $\vec{e} = [e_1, e_2, ..., e_T]$ y otro de valores logarítmicos del pitch $\vec{p} = [p_1, p_2, ..., p_T]$ donde T es el número de ventanas de la locución de voz. La eliminación de los segmentos que no son voz se consigue mediante el uso de un Detector de Actividad Vocal (VAD), aceptando únicamente aquellas ventanas con valor de pitch y energía mayores que un umbral θ. El umbral elegido θ es:

$$\theta = min\{\vec{e}\} + \frac{MD}{10}$$

donde MD es el Margen Dinámico de la energía, $MD = max\{\vec{e}\} - min\{\vec{e}\}$

Para obtener información de la velocidad de los vectores de energías $\vec{e}$ y pitch $\vec{p}$, los valores Δ se obtienen como la diferencia entre ventanas consecutivas. Así, $\Delta_{e_k} = e_{k+1} - e_k$.

Como refleja la Figura 19, por cada locución de voz u, la parametrización prosódica consiste en un conjunto de $d = 4$ vectores de características o tramas (energía, pitch y sus valores Δ).

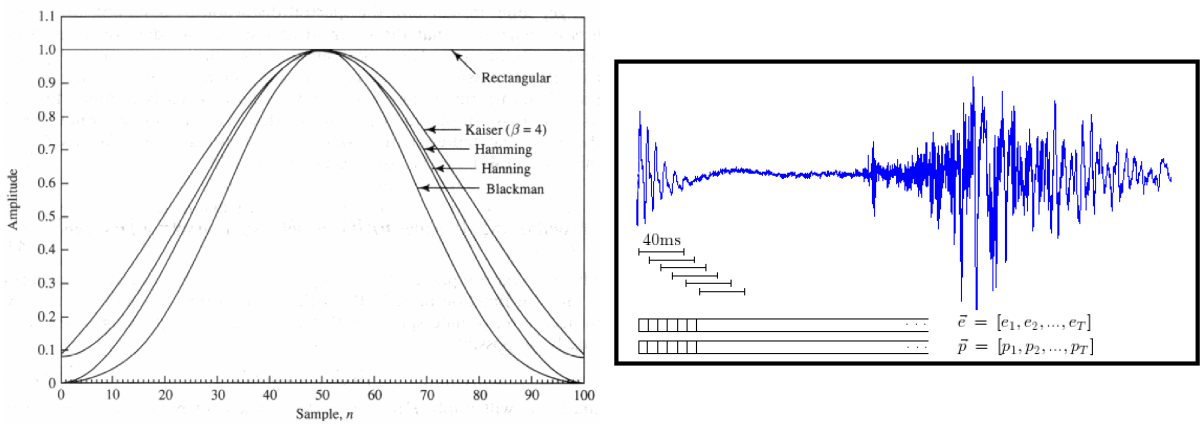

Figura 18: a) Ventanas temporales más utilizadas para el enventanado de la señal de voz. b) Enventanado y vectores de energía $\vec{e}$ y pitch $\vec{p}$ de la señal de voz.

$$\vec{u_p} = \{\vec{e}, \vec{p}, \vec{\Delta_e}, \vec{\Delta_p}\}$$

Es posible normalizar cada una de las 4 tramas restándole su valor medio. En el capítulo 5 de Pruebas y Resultados se indicará que tipo de normalización se ha llevado a cabo según el sistema o el tipo de prueba realizada, para optimizar resultados.

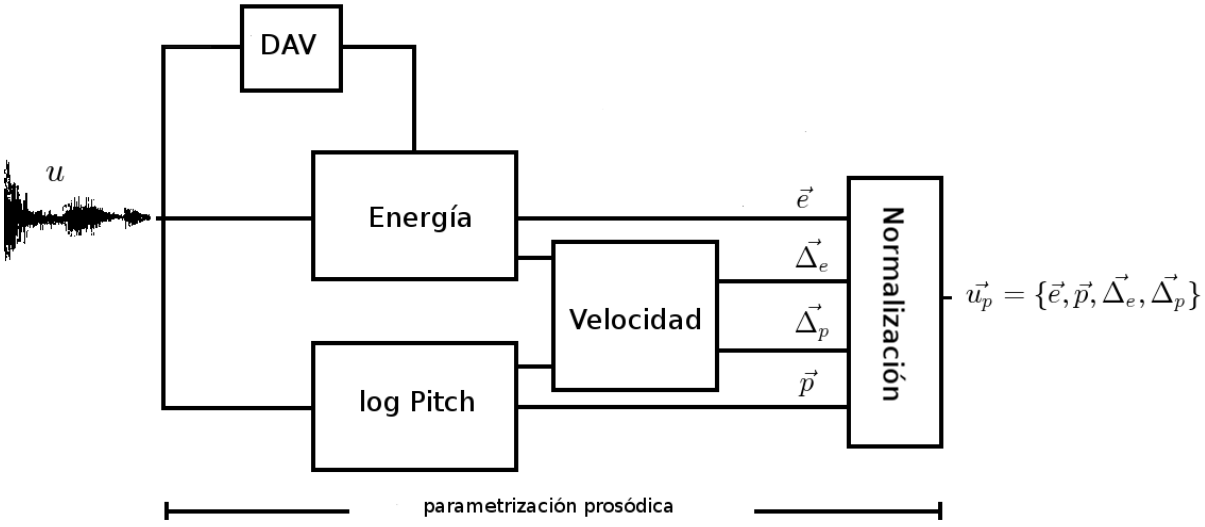

Figura 19: Diagrama de bloques de la extracción de parámetros prosódicos de la señal de voz.

4.2.2. Subsistemas front-end (SVM con estadísticos y GMM-SVM)

Un sistema de reconocimiento de voz *front-end* es todo aquel que utiliza como entrada la propia señal de voz y obtiene a la salida una serie de puntuaciones de similitud de dicha señal de voz frente a un conjunto de modelos previamente entrenados.

Para la tarea que nos ocupa se han diseñado dos subsistemas *front-end*.

• Un sistema de SVM cuyo vector de entrada es un conjunto de estadísticos globales de las características prosódicas.
• Otro sistema de SVM que utiliza los valores de las medias de los GMMs para configurar el supervector de entrada.

A partir de ahora al primero le llamaremos **SVM con estadísticos** y al segundo **GMM-SVM**.

En la siguiente sección se describen los procesos de modelado y evaluación de los subsistemas y la fusión de los resultados obtenidos por ambos.

Creación y evaluación de los modelos del subsistema SVM basado en estadísticos globales

Este tipo de modelado SVM utiliza como vector de entrada un vector formado por la concatenación de $n = 9$ valores estadísticos de cada uno de las $d = 4$ tramas prosódicas ($\vec{e}$, $\vec{\Delta}_e$, $\vec{p}$ y $\vec{\Delta}_p$). Estos 9 coeficientes estadísticos aparecen en la Tabla 3.

Coeficientes
Máximo
Mínimo
Medio
Desviación estándar
Mediana
Primer cuartil
Tercer cuartil
Skewness
Kurtosis

Tabla 3: Coeficientes estadísticos calculados por cada trama prosódica.

Por lo tanto, por cada locución de voz se obtiene un vector de longitud fija de $L = d * n = 4 * 9 = 36$ valores. En este nuevo espacio de características L-dimensional es donde se modelan las emociones usando un SVM lineal. Como puede verse en la Figura 20 el vector de rasgos L-dimensional se puede ver como el resultado de la función *kernel* [37] $l(\vec{u}_p)$ que mapea las tramas prosódicas de $\vec{u}_p$ en un espacio de características L-dimensional.

Con los datos de entrenamiento se crean los modelos por cada emoción. Dado un modelo SVM $\vec{w}_e$ de una emoción e, la función de puntuación o *scoring* $s(\vec{w}, l(\vec{u}))$ por cada locución de test $\vec{u}_{p_{test}}$ es simplemente un producto escalar calculado de la siguiente forma [Ver Figura 20]:

$$s(\vec{w}_e, l(\vec{u}_{p_{test}})) = \vec{w}_e * l(\vec{u}_{p_{test}})^T$$

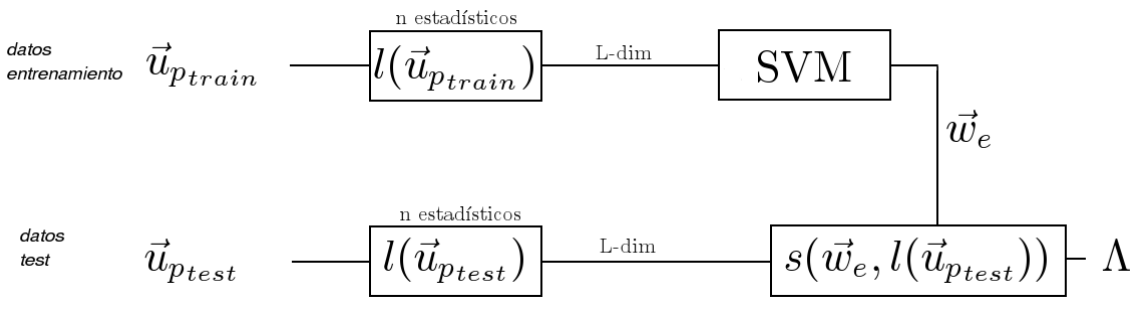

Figura 20: Diagrama de bloques del clasificador SVM utilizando estadísticos globales.

Como resultado de dicha función de *scoring* se tienen una puntuación Λ que dará una medida de la similitud entre la locución de test $\vec{u}_{p_{test}}$ y el modelo $\vec{w}_e$.

La Figura 21 representa un esquema del funcionamiento de un SVM desde el punto de vista de la distribución de los datos de entrenamiento. Por cada emoción $e1$, $e2$ y $e3$ existen N_{e_1}, N_{e_2} y N_{e_3} locuciones de entrenamiento respectivamente. Así, para entrenar el modelo $\vec{w}_{e_1}$ se usan como datos *target* (clase +1) las locuciones $l(\vec{u}_{p_{train}})$ pertenecientes a la emoción $e1$ y como datos *non-target* (clase -1) a los pertenecientes al resto de clases o emociones, en el ejemplo a las emociones $e2$ y $e3$.

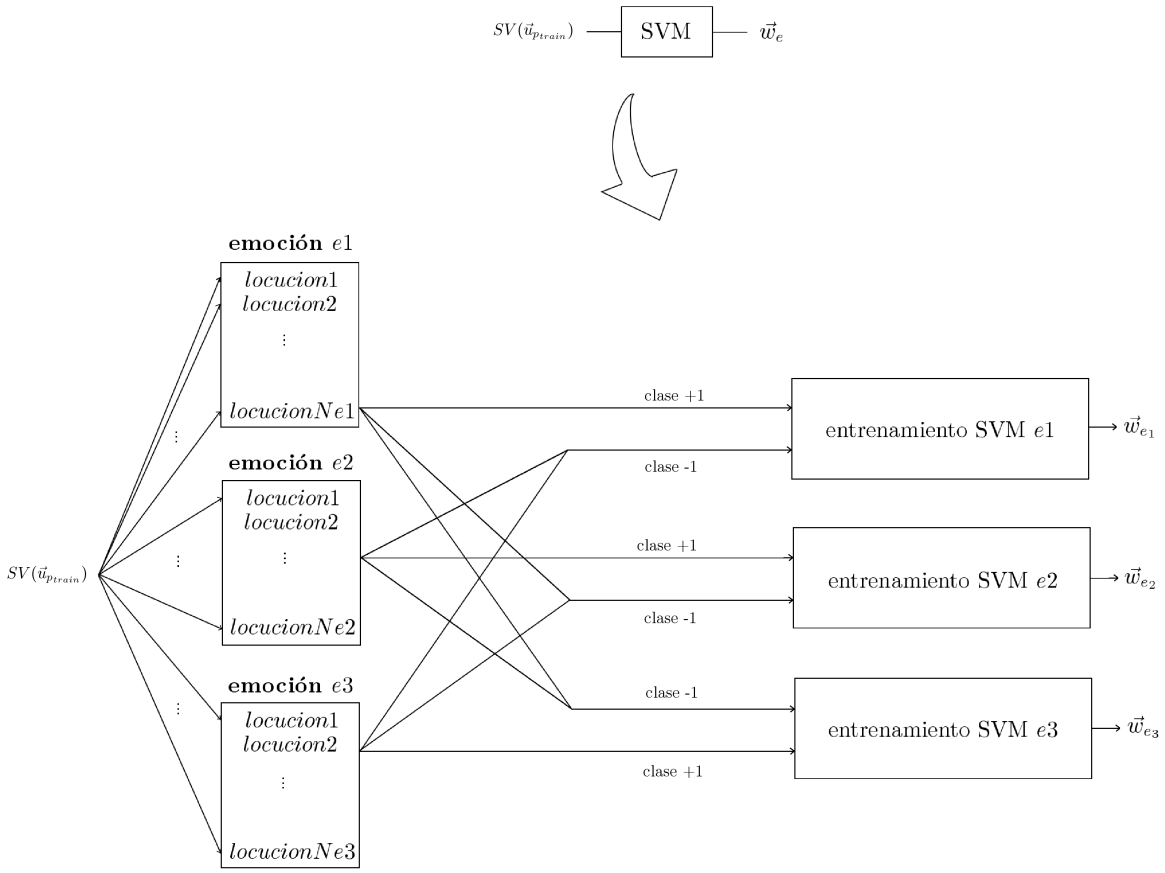

Figura 21: Esquema de distribución de los datos de entrenamiento en un clasificador SVM para vectores de entrada $l(\vec{u}_{p_{train}})$.

Creación y evaluación de los modelos del subsistema GMM-SVM

Como ya se ha explicado en el capítulo de Técnicas de Reconocimiento de Emociones [3.7], la técnica de SVMs basados en supervectores GMMs consiste en entrenar los modelos SVM con supervectores L$'$-dimensionales creados mediante la apilación de los vectores medios d-dimensionales de las M componentes Gaussianas, donde $L' = M * d$.

Se puede considerar al supervector GMM como resultado de una función $SV(\vec{u}_p)$ que mapea los vectores prosódicos $\vec{u}_p$ en un vector de mayor dimensión $L' = M * d$ [Ver Figura 22]. En este espacio L'-dimensional es donde se modela el SVM para obtener un modelo final $\vec{w}_e$ de la emoción e.

En nuestro caso la parametrización prosódica $\vec{u}_p$ consiste en 4 vectores $(\vec{e}, \vec{p}, \vec{\Delta}_e, \vec{\Delta}_p)$ por lo

tanto el vector medio de cada mezcla GMM serán 4-dimensionales [Ver Figura 22]. Tomando por ejemplo un número de Gaussianas de 256 ($M = 256$), el supervector GMM $SV(\vec{u}_p)$ que servirá como entrada al SVM tendrá una dimensión de $L^{'} = 256 * 4 = 1024$.

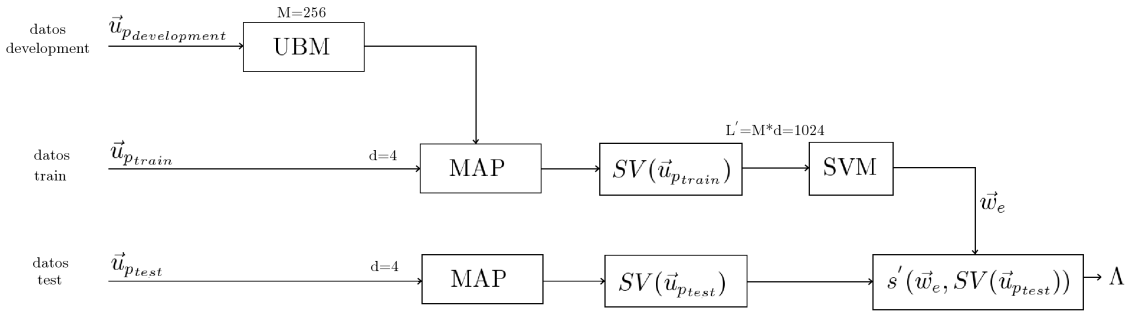

Figura 22: Diagrama de bloques del clasificador GMM-SVM.

Para este caso la función de *scoring* también consistirá en un producto escalar entre el modelo $\vec{w}_e$ y el supervector GMM de test $SV(\vec{u}_{p_{test}})$ siendo ésta:

$$s^{'}(\vec{w}_e, SV(\vec{u}_{p_{test}})) = \vec{w}_e * SV(\vec{u}_{p_{test}})^T$$

La manera en que el SVM funciona a la hora de clasificar aparece en la Figura 23. El funcionamiento es el mismo al de la Figura 21 excepto por el hecho de que los vectores de entradas del SVM son supervectores GMM de la forma $SV(\vec{u}_{p_{train}})$.

Fusión suma de los resultados de los subsistemas

Tanto el sistema de SVMs con estadísticos globales como el de supervectores GMM ofrecen a la salida unas puntuaciones de similitud entre la muestra de test $\vec{u}_{p_{test}}$ y el modelo $\vec{w}_e$:

$$s(\vec{w}_e, l(\vec{u}_{p_{test}})) = \vec{w}_e * l(\vec{u}_{p_{test}})^T$$
$$s^{'}(\vec{w}_e, SV(\vec{u}_{p_{test}})) = \vec{w}_e * SV(\vec{u}_{p_{test}})^T$$

respectivamente.

Mediante la combinación de dichas puntuaciones se consigue una nueva puntuación final $S(\vec{w}_e, \vec{u}_{p_{test}}))$ que puede ofrecer mejores resultados si los subsistemas fusionados dan información complementaria.

La combinación se realiza como una fusión suma (*sum fusion* en inglés) precedida de una *T-norm* (*test normalization*) [capítulo 2.5] que hace que los rangos de puntuaciones de ambos sistemas sean similares. El conjunto *cohorte* de la *T-norm* está formado por todo el conjunto de emociones $\vec{w}_e$ para $e = 1, ..., N_{emociones}$. La puntuación final fusionada $S(\vec{w}_e, \vec{u}_{p_{test}}))$ se calcula como:

$$S(\vec{w}_e, \vec{u}_{p_{test}}) = \frac{s'(\vec{w}_e, SV(\vec{u}_{p_{test}})) - \mu'}{std'} + \frac{s(\vec{w}_e, l(\vec{u}_{p_{test}})) - \mu}{std}$$

Donde μ' y μ son las medias de las puntuaciones *cohorte*, y std' y std son sus respectivas desviaciones estándares.

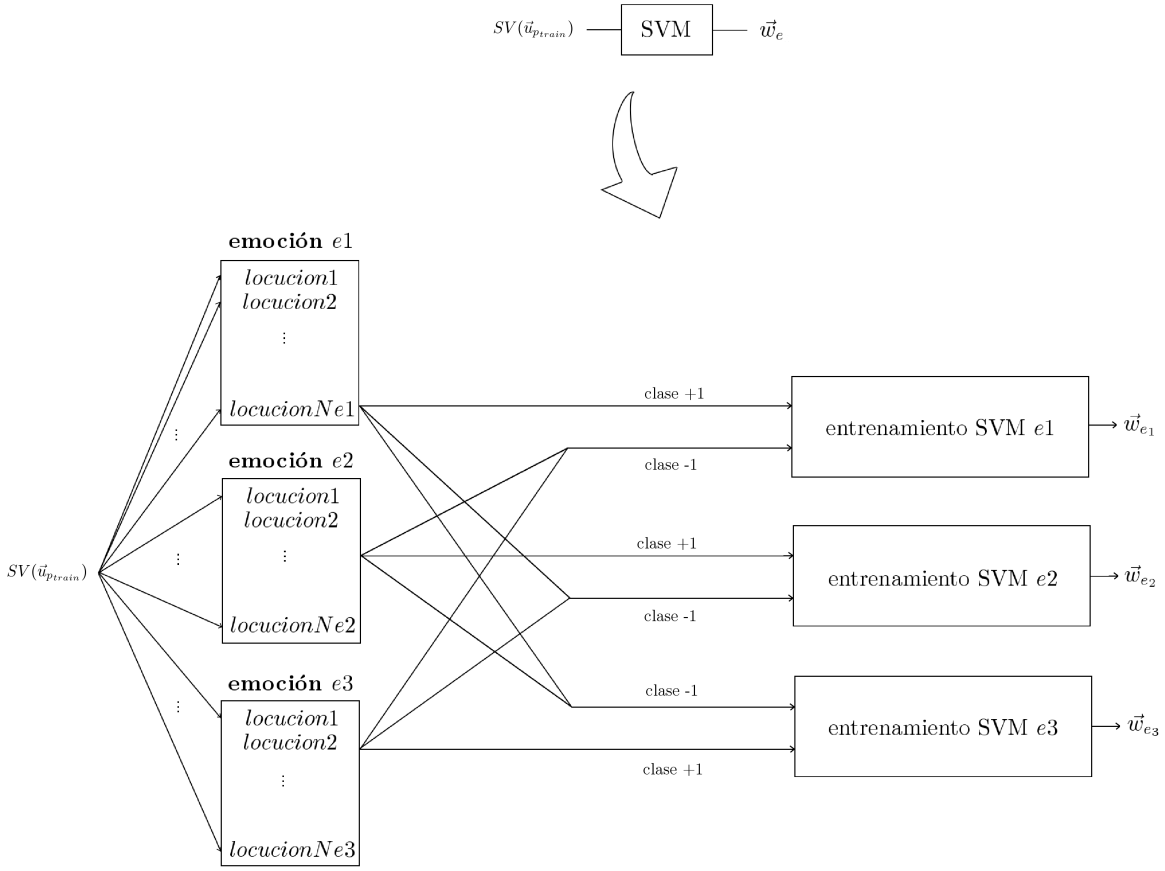

Figura 23: Esquema de distribución de los datos de entrenamiento en un clasificador SVM para supervectores de entrada $SV(\vec{u}_{p_{train}})$.

4.2.3. Sistema back-end (Fusion *Anchor Models*)

Por contraposición a los sistemas *front-end*, tenemos los sistemas *back-end*. Como ejemplo de este tipo de técnicas, tenemos la fusión de los *anchor models* (*Anchor Model Fusion*, AMF) que ya vimos en el capítulo 3.7.4.

Esta técnica novedosa es original del ATVS y fue presentada en el congreso internacional *Interspeech 2008* para reconocimiento automático de idioma [27]. La aplicación a tareas de reconocimiento de emociones la presentamos en [28] y está aceptada y pendiente de presentar en *Interspeech 2009*.

Creación de modelos SVM a partir de los resultados de los subsistemas *front-end*

Este tipo de técnicas usa las puntuaciones obtenidas previamente por otros subsistemas y las utiliza para formar el vector de parámetros de entrada para el nuevo sistema que producirá las puntuaciones finales. En la Figura 15 se vio como una locución de test $\vec{x}$ se enfrentaba a n modelos de N_{sist} subsistemas para así conformar el vector de puntuaciones de dimensión $N = n * N_{sist}$ denotado como $\vec{S}_{x,m}$. Este vector de puntuaciones es el que pasa a ser el vector de parámetros de la locución $\vec{x}$ para el sistema *back-end*.

En nuestro caso, el número de subsistemas N_{sist} es de 2, el sistema GMM-SVM y el de SVM con estadísticos. Por otro lado, el número de modelos n a enfrentar dependerá de la base de datos que usemos. Así, por ejemplo para *SUSAS Simulated* en un sistema independiente de locutor tendremos 11 modelos, uno por cada emoción y por lo tanto $n = 11$. [Figura 24]

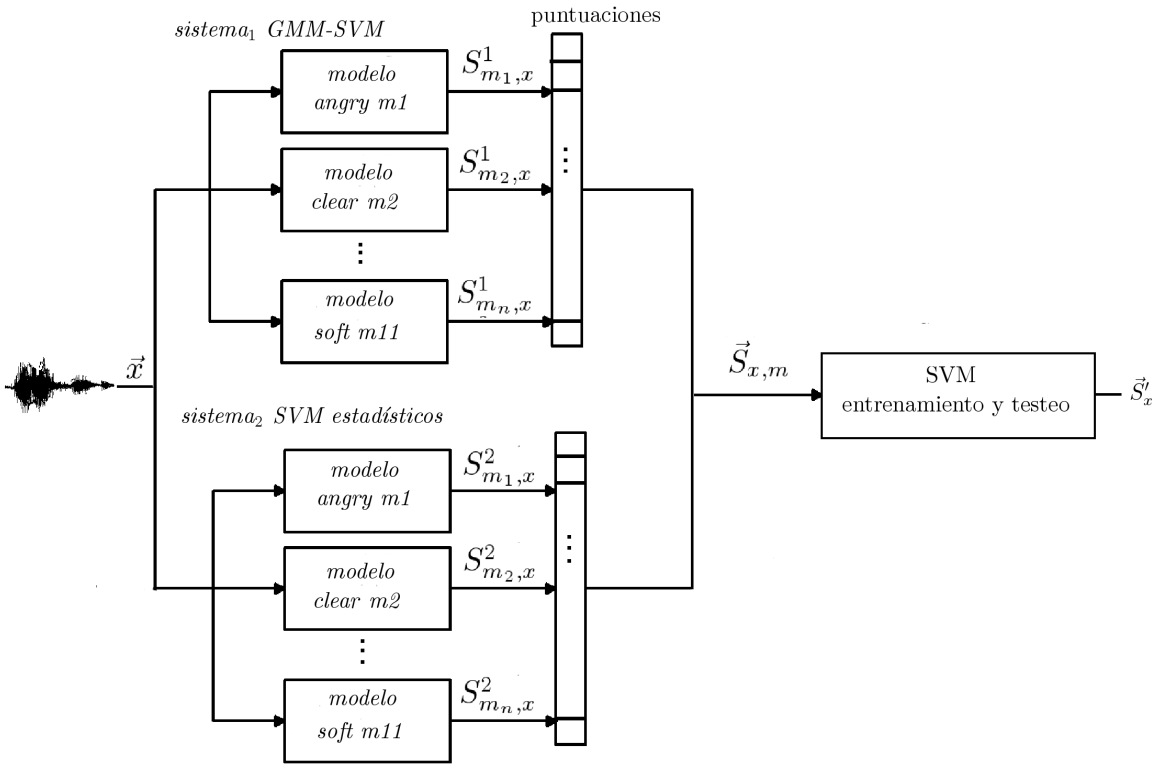

Figura 24: Uso de las puntuaciones de dos sistemas *front-end* para conformar el sistema *back-end* para la base de datos *SUSAS Simulated*.

Una vez se ha obtenido el nuevo vector de parámetros $\vec{S}_{x,m}$ para cada locución $\vec{x}$, el siguiente paso es entrenar un clasificador *back-end* con esta nueva parametrización. El nuevo clasificador *back-end* va a ser un SVM. El entrenamiento de los modelos SVM se hará de la misma forma que vimos en la sección 4.2.2. Los datos de entrenamiento (pertenecientes al espacio de *anchor models*) serán los encargados de modelas los nuevos modelos w'_e y los datos de test (también pertenecientes al espacio de *anchor models*) los evaluarán obteniendo una puntuación final $\vec{S}'_x$ [Ver Figura 24].

5
Pruebas y Resultados

En este trabajo se distiguen dos tipos de experimentos: independientes y dependientes de locutor. Se lleva a cabo dicha división con el objetivo de analizar la variabilidad introducida por los distintos usuarios.

En aplicaciones donde no existen datos específicos por cada locutor es preferible usar sistemas independientes de locutor. Mientras que si sí están disponibles datos de cada locutor es mejor adaptar los modelos a cada uno de ellos eliminando la variabilidad inter-locutor y así presumiblemente conseguiremos reducir la tasar de error.

La ventaja de los sistemas independientes de locutor es que no es necesario el entrenamiento de modelos específicos para cada usuario. Por ello, existe un compromiso entre ambos tipos de sistemas. Los independientes de locutor ofrecen una mayor rapidez y comodidad para el usuario mientras que los dependientes de locutor consiguen menores tasas de error.

Para cada uno de estos dos tipos de experimentos y para cada base de datos se van a presentar y analizar los resultados obtenidos mediante los dos subsistemas *front-end* y su fusión suma, al igual que para el sistema *back-end* de AMF.

Con el objetivo de lograr sistemas más robustos se ajustarán una serie de variables como el *coste* asociado al entrenamiento SVM o el número de mezclas Gaussianas de los GMM, además de la normalización de tanto los vectores de parámetros prosódicos como de las puntuaciones resultantes.

5.1. Pruebas y Resultados independientes de locutor

Para evaluar los sistemas independientes de locutor se han hecho dos tipos de pruebas.

En las primeras, se evalúan los modelos de cada base de datos (*SUSAS Simulated*, *SUSAS Actual* y *Ah3R1*) frente a datos de test de la misma base de datos. Por ejemplo, las locuciones de test de *Ah3R1* se evaluarán únicamente frente a los modelos creados a partir de la base de datos *Ah3R1*. A este tipo de experimentos los llamaremos experimentos *Intra*-Base de datos.

En el otro tipo de pruebas se evalúan las locuciones de test de cada base de datos frente a todos los modelos creados por todas las bases de datos. Es decir, por ejemplo, los datos de test de *Ah3R1* se evalúan frente a los modelos de *SUSAS Simulated*, *SUSAS Actual* y *Ah3R1*. Serán llamados por lo tanto experimentos *Inter*-Base de datos.

5.1.1. Experimentos *Intra*-Base de datos: Evaluación de cada Base de Datos frente a modelos de la misma Base de Datos

SUSAS Simulated

En este apartado se van a describir los experimentos independientes de locutor realizados sobre la base de datos *SUSAS Simulated*. Como se vio en el capítulo 4.1.1, se tiene 9 locutores los cuales se dividen en 3 grupos según la etapa (*development*, entrenamiento y test) a la que se dediquen [Ver Tabla 4]. Los datos de *development* serán utilizados para generar el modelo UBM.

Etapa	Locutores
Development	*g1,b1,n1*
Entrenamiento	*g2,b2,n2*
Test	*g3,b3,n3*

Tabla 4: Distribución de locutores para experimentos independientes de locutor en *SUSAS Simulated*.

- ***SUSAS Simulated* - SVM con estadísticos**

Como aparece en la Figura 25, se entrenan 11 modelos ($\vec{w}_{SVM\_angry}$, $\vec{w}_{SVM\_clear}$, $\vec{w}_{SVM\_cond50}$, $\vec{w}_{SVM\_cond70}$, $\vec{w}_{SVM\_fast}$, $\vec{w}_{SVM\_lombard}$, $\vec{w}_{SVM\_loud}$, $\vec{w}_{SVM\_neutral}$, $\vec{w}_{SVM\_question}$, $\vec{w}_{SVM\_slow}$ y $\vec{w}_{SVM\_soft}$), uno por cada emoción utilizando los locutores de entrenamiento (*g2, b2, n2*). El número de locuciones de entrenamiento por cada emoción es de:

35 palabras * 2 repeticiones/palabra * 3 locutores = 210 locuciones/emoción.

Para este caso en que no interviene la técnica GMM, no se entrena un UBM y por lo tanto no se usan los datos de los locutores *g1, b1, n1*.

Una vez se tiene un modelo por cada emoción se pasa a la etapa de evaluación de los mismos. Se usan los datos de test de los locutores *g3, b3, n3*. Se enfrentan todos las locuciones de test frente a los 11 modelos.

El número de locuciones de test es de:

35 palabras * 2 repeticiones/palabra * 3 locutores * 11 emociones = 2310 locuciones.

Por lo tanto, como cada locución de test se enfrenta a los 11 modelos, el número de puntuaciones será de:

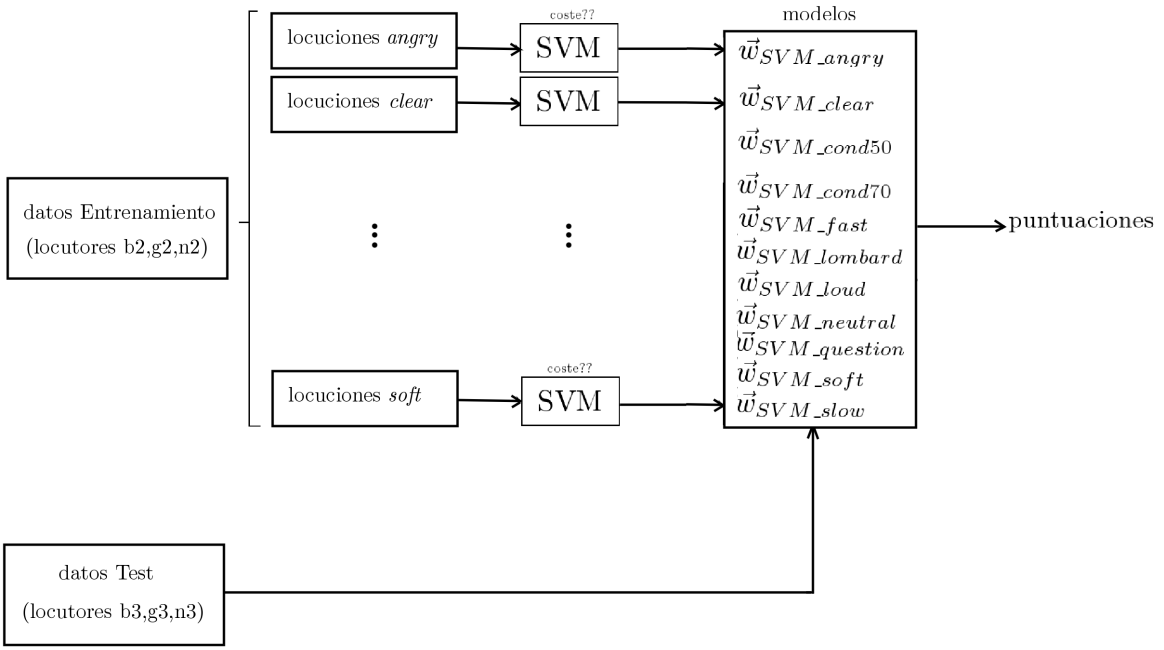

Figura 25: Esquema de las pruebas independientes de locutor para el sistema '*SUSAS Simulated* - SVM con estadísticos'.

11 modelos * 2310 locuciones = 25410 puntuaciones.

Para este subsistema se van a llevar a cabo las siguientes tareas para optimizar los resultados:

- Normalización de los vectores de parámetros prosódicos
- Optimización variable *coste* de entrenamiento
- Selección de parámetros
- T-normalización de puntuaciones

Como se comentó en el capítulo 4.2.1, es posible realizar una normalización de cada una de las 4 tramas de vectores prosódicos ($\vec{e}$, $\vec{p}$, $\vec{\Delta}_e$ y $\vec{\Delta}_p$) restándole su valor medio. Para estos experimentos se ha realizado la normalización del vector $\vec{e}$ pues es la opción que mejores resultados consigue. Consecuentemente los vectores prosódicos son:

$$\vec{u_p} = \{\vec{e} - E(\vec{e}), \vec{p}, \vec{\Delta_e}, \vec{\Delta_p}\}$$

donde $E(\vec{e})$ es la esperanza matemática o valor medio del vector de energías $\vec{e}$.

Otra de las variables que se van a ajustar es el *coste* del clasificador SVM. El coste en el entrenamiento SVM (ver sección 3.7.2) es una variable mediante la cual controlamos la penalización aplicada a una muestra incorrectamente clasificada a la hora de establecer el hiperplano de separación entre las clases.

Los resultados para varios valores de *coste* se muestran en la Figura 26 en forma de curva DET y en la Tabla 5 con valores numéricos.

Una conclusión que se puede sacar aunque no se refleja en los resultados anteriores, es que cuanto mayor es el *coste*, mayor tiempo se emplea en el entrenamiento de los modelos. Por esa razón interesa el menor valor de *coste* posible. Según se ve en los resultados, éstos son mejores con un valor de *coste* de 10. Por lo tanto, y dado que dicho valor no hace que el tiempo de entrenamiento se dispare, se eligirá 10 como valor de *coste*.

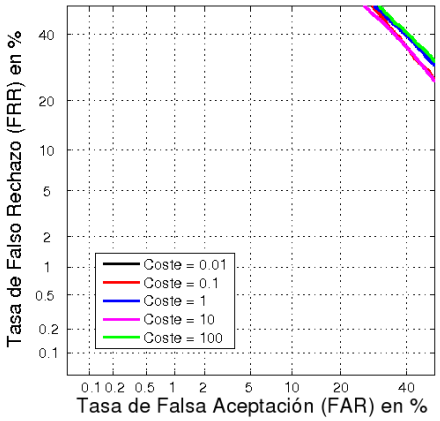

Figura 26: Curvas DET del sistema '*SUSAS Simulated* - SVM con estadísticos' para diferentes costes de entrenamiento.

Coste	**EER global(%)**	**DCF**$_{min}$	**EER medio (%)**
0.01	39.85	0.099	37.20
0.1	38.18	0.098	36.11
1	39.85	0.099	37.20
10	38.07	0.095	35.74
100	40.40	0.098	36.52

Tabla 5: Resultados '*SUSAS Simulated* - SVM con estadísticos' dependiendo del valor de la variable *coste* de entrenamiento.

El último tipo de optimización que se ha realizado sobre este tipo de experimentos es la selección de los mejores coeficientes estadísticos de la Tabla 3, eliminando aquellas que ofrecen información redundante. La técnica usada para la selección es *backward-elimination* que consiste en a partir de todos los parámetros ir secuencialmente eliminando áquel que más decrementa o menos incrementa el porcentaje de clasificación.

El proceso de selección de características *backward-elimination* nos ha llevado a concluir que la mejor configuración se obtiene eliminando el coeficiente de *kurtosis*, la mediana y la media del vector de energías $\vec{e}$.

Una vez llevada a cabo la selección de características y tras hacer T-normalización de los resultados, llegamos a obtener los resultados de la Tabla 6:

Norm. $\vec{u}_p$	**Coste**	***Backward elimination***	**T-norm**	**EER**$_{global}$	**DCF**$_{min}$	**EER**$_{medio}$
$\vec{e}$	10	kurtosis mediana y media de $\vec{e}$	sí	35.11	0.096	34.47

Tabla 6: Configuración y resultados optimizados para '*SUSAS Simulated* - SVM con estadísticos'.

- ***SUSAS Simulated* - GMM-SVM**

Para la técnica de GMM-SVM, se usan los datos de *development* (g1, b1, n1) para entrenar el modelo UBM que nos servirá como base para la adaptación a los modelos GMM. El número

de datos de *development* es de:

35 palabras * 2 repeticiones/palabra * 3 locutores * 11emociones = 2310 locuciones.

Los datos de entrenamiento (g2,b2,n2) adaptaran dicho UBM generando así un modelo GMM por cada locución. [Ver Figura 27]

Como ya se explicó en el capítulo 3.7.3, por cada locución de entrenamiento y test se concatenan los vectores de medias 4-dimensionales de las M componentes gaussianas conformando así el supervector de entrada al clasificador SVM. El valor M será ajustado para obtener los mejores resultados. Como se aprecia en la Figura 27, mediante los clasificadores SVM se entrenan 11 modelos, uno por emoción ($\vec{w}_{GMM-SVM\_angry}$, $\vec{w}_{GMM-SVM\_clear}$, $\vec{w}_{GMM-SVM\_cond50}$, $\vec{w}_{GMM-SVM\_cond70}$, $\vec{w}_{GMM-SVM\_fast}$, $\vec{w}_{GMM-SVM\_lombard}$, $\vec{w}_{GMM-SVM\_loud}$, $\vec{w}_{GMM-SVM\_neutral}$, $\vec{w}_{GMM-SVM\_question}$, $\vec{w}_{GMM-SVM\_slow}$ y $\vec{w}_{GMM-SVM\_soft}$). Al igual que el sistema de SVM con coeficientes estadísticos, se dispone de 210 locuciones de entrenamiento por emoción.

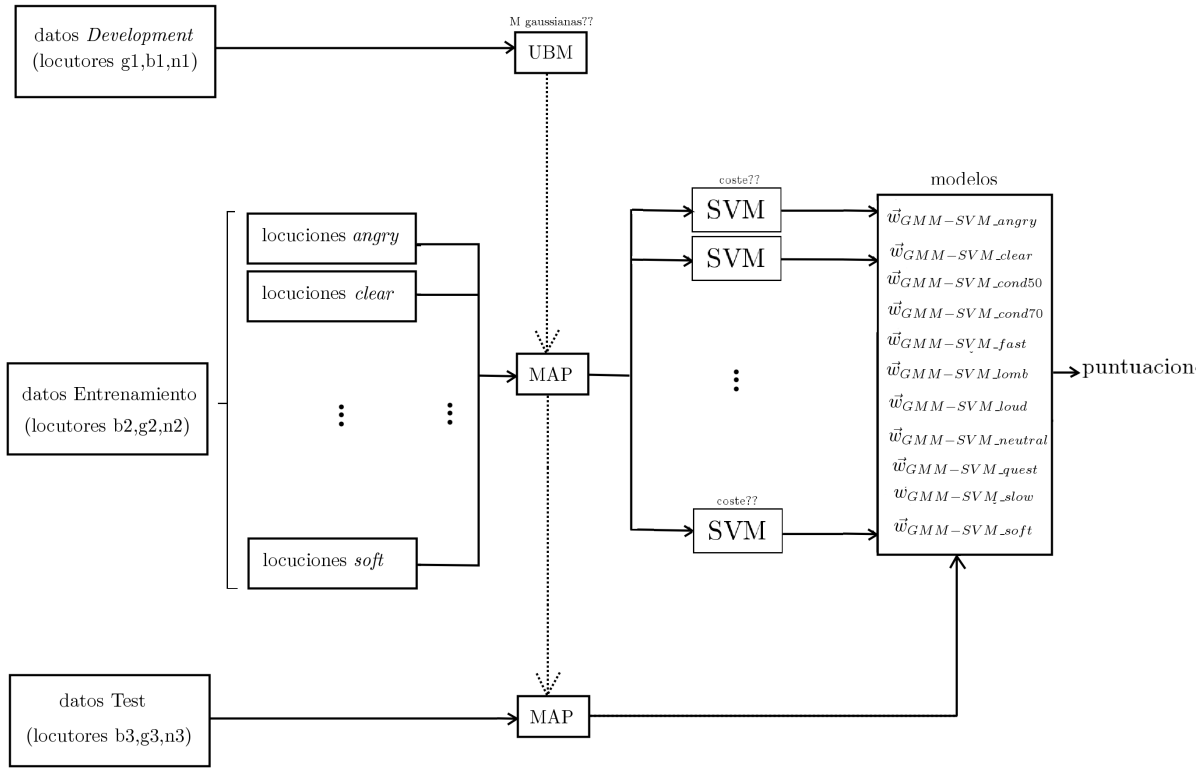

Figura 27: Esquema de las pruebas independientes de locutor para '*SUSAS Simulated* - GMM-SVM'.

En este sistema las tareas que se van a realizar para optimizar resultados son:

- Normalización de los vectores de parámetros prosódicos
- Optimización variable M número de gaussianas
- Optimización variable coste de entrenamiento
- T-normalización de puntuaciones

Tras una serie de pruebas realizadas normalizando cada uno de los vectores de parámetros prosódicos de $\vec{u_p}$ se ha llegado a la conclusión que la configuración que ofrece menor tasa de error es mediante la normalización de tanto el vector de energías $\vec{e}$ como el de su velocidad $\vec{\Delta}_e$

quedando la parametrización prosódica de la siguiente manera:

$$\vec{u_p} = \{\vec{e} - E(\vec{e}), \vec{p}, \vec{\Delta_e} - E(\vec{\Delta_e}), \vec{\Delta_p}\}$$

El siguiente valor a ajustar es M, el número de componentes gaussianas de los GMM. La ventaja de modelar con un número alto de gaussianas es que se logra una mejor adaptación de las mezclas a los datos del problema. La desventaja es que se necesita disponer de gran cantidad de datos. Para un valor de M bajo se produce una peor adaptación al problema pero por el contrario no requiere de muchos datos.

Los resultados de esta optimización se muestran en la Figura 28 en forma de curva DET y en la Tabla 7 mediante valores numéricos.

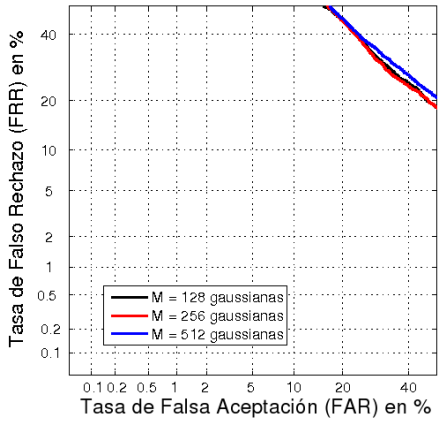

Figura 28: Curvas DET del sistema '*SUSAS Simulated* - GMM-SVM' para varios números de Gaussianas.

M	EER global(%)	DCF_{min}	EER medio (%)
128	31.09	0.0901	29.80
256	30.83	0.0903	29.62
512	32.43	0.0911	31.98

Tabla 7: Resultados para '*SUSAS Simulated* - GMM-SVM' dependiendo del número de gaussianas M.

Analizando los resultados de la Tabla 7 se opta por un valor de M de 256 gaussianas pues aunque no es la que mejor DCF_{min} ofrece, sí es la que menor tasa de error consigue, tanto global como media.

A continuación se ajustará la variable *coste* manteniendo fijo el número de gaussianas a 256. La Figura 29 refleja los resultados para distintos valores de *coste* de entrenamiento, mientras que la Tabla 8 los muestra numéricamente.

Como vemos en la Tabla 8 hay discordancia entre el valor de *coste* que hace optimizar cada una de las 3 medidas de resultados. Aunque con un coste de 100 se obtiene el mejor resultado de DCF_{min}, no se optará por dicha opción pues necesita un tiempo de entrenamiento mayor. Con un coste de 0.01 se obtiene la mejor tasa de EER_{medio}. Sin embargo, este coste tampoco será escogido. La mejor opción es tomar un *coste* de 1. De esta manera únicamente se empeora

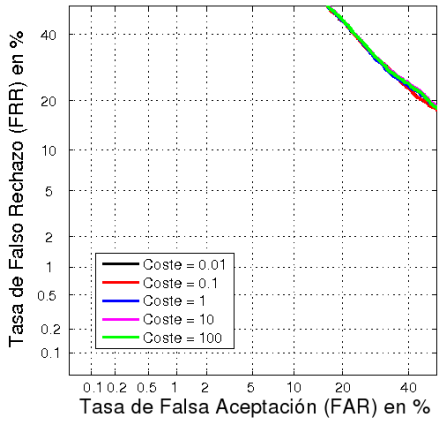

Figura 29: Curvas DET para varios valores de *coste* en '*SUSAS Simulated* - GMM-SVM'.

Coste	EER global(%)	DCF_{min}	EER medio (%)
0.01	31.05	0.0904	29.60
0.1	31.02	0.0903	29.63
1	30.83	0.0903	29.62
10	31	0.0902	29.84
100	30.92	0.0901	29.87

Tabla 8: Resultados dependiendo del *coste* para '*SUSAS Simulated* - GMM-SVM'.

2 centésimas el EER_{medio} con respecto al de coste 0.01 y la EER_{global} se ve mejorada en casi 2 décimas.

Tras haber optimizado tanto el valor de *coste* como el de M, la última tarea es realizar una T-normalización de los resultados utilizando la configuración de la Tabla 9.

Norm. $\vec{u}_p$	M	coste	*T-norm*	EER_{global}	DCF_{min}	EER_{medio}
$\vec{e}$ y $\vec{\Delta}_e$	256	10	sí	29.44	0.0903	30.44

Tabla 9: Configuración y resultados optimizados para '*SUSAS Simulated* - GMM-SVM'.

- ***SUSAS Simulated* - Fusión suma SVM estadísticos + GMM-SVM**

El capítulo 4.2.2 describió en que consistía la fusión suma. Dicha fusión se ha de realizar previa T-normalización de las puntuaciones para que los rangos de puntuaciones de tanto el subsistema GMM-SVM como el de SVM con estadísticos sean similares.

A la hora de realizar la fusión se toma para cada uno de los dos sistemas la configuración que ofrece mejores resultados [Tabla 6 y 9]. En la Figura 30 se representa la curva DET para cada sistema y para la fusión de ambos.

Los valores de EER global de GMM-SVM, SVM con estadísticos y la fusión suma son 29.44 %, 35.11 % y 31.62 % respectivamente. Para este caso la fusión suma no consigue mejorar los resultados del mejor de los dos subsistemas pues el otro obtiene resultados bastante peores.

- ***SUSAS Simulated* - Fusión de *Anchor Models (AMF)***

Como vimos en el capítulo 4.2.3, para esta nueva técnica se utilizan las puntuaciones de cada

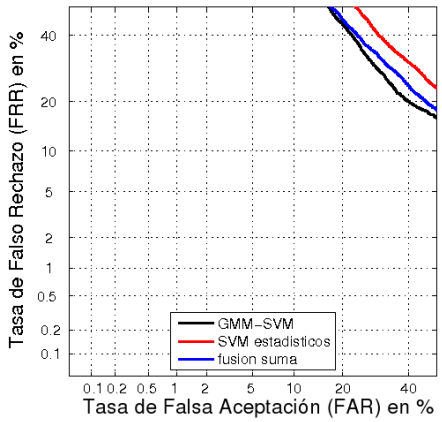

Figura 30: Curvas DET de '*SUSAS Simulated* - SVM con estadísticos, GMM-SVM y fusión suma'.

locución de test obtenidas tras evaluarla frente a los 11 modelos de cada uno de los subsistemas GMM-SVM y SVM con estadísticos para conformar un nuevo vector de parámetros. Dichas puntuaciones serán las correspondientes a la configuración que en cada caso ha dado los mejores resultados [Tabla 6 y 9]. Dicho vector $\vec{S}_{locucion\_test}$ tendrá 22 valores [Ver Figura 31].

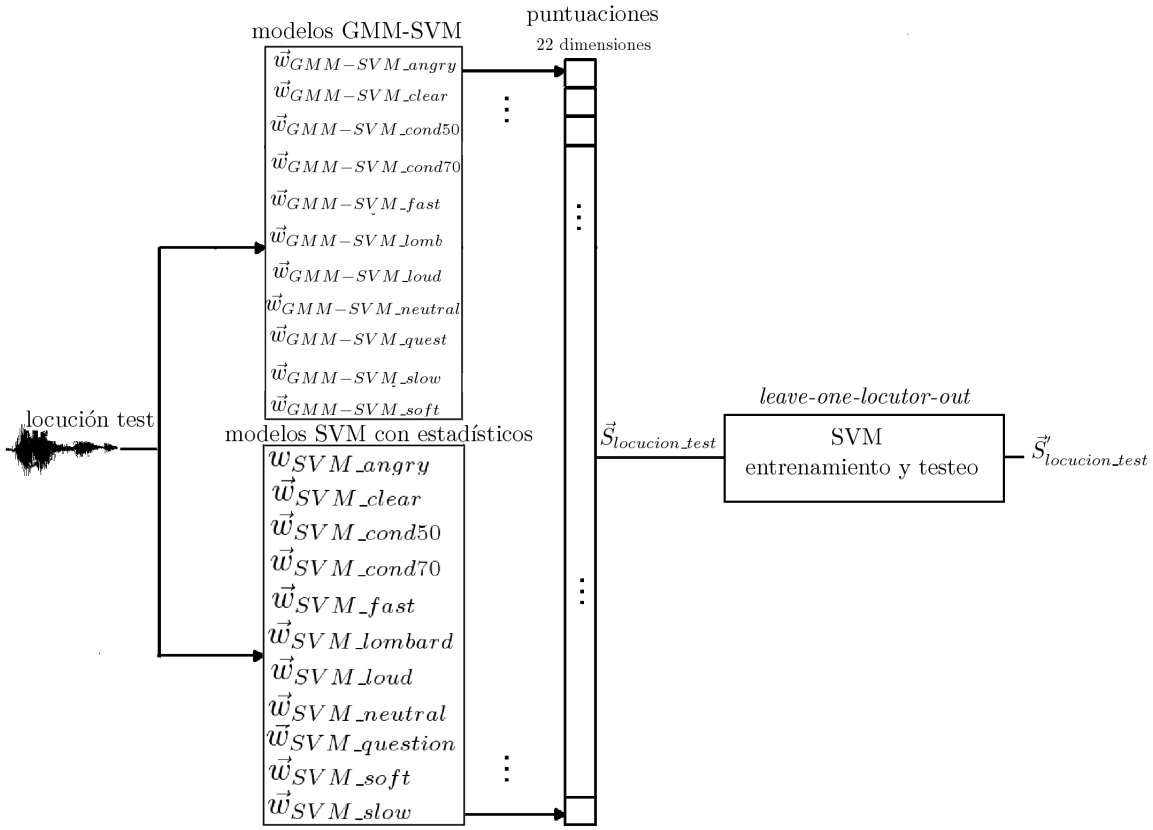

Figura 31: Esquema de las pruebas independientes de locutor para '*SUSAS Simulated* - AMF'.

Una vez se tiene por cada locución de test $\vec{x}$ un nuevo vector de parámetros $\vec{S}_x$, éstos se utilizan como entrada a un clasificador SVM. Para mantener los experimentos independientes de locutor, se cogerán iterativamente los datos de cada uno de los tres locutores *g3,b3,n3* y se utilizarán para evaluación mientras que los datos de los otros dos restantes se utilizarán para entrenar los modelos SVM, uno por emoción. A esta práctica se la conoce como *leave-one-*

locutor-out que es un tipo de validación cruzada (*cross-validation*). Así se consigue que datos de un mismo locutor no se utilicen para entrenamiento y test simultáneamente. A esta técnica se la conoce como validación cruzada (*cross validation*) [38].

En la Figura 32 se representan un conjunto de curvas DET para varios valores de la variable *coste* del clasificador *back-end* SVM. Y en la Tabla 10 valores numéricos de tasas de error y DCF_{min}. La mejor configuración se logra cuando el *coste* toma valor 1. Aunque para un coste de 10 se reduce en 2 milésimas el DCF_{min}, ésto supone un mayor tiempo de entrenamiento y peores tasas de error.

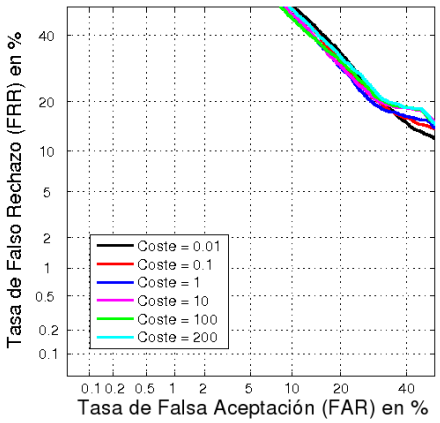

Figura 32: Curvas DET de '*SUSAS Simulated* - AMF' para varios valores de *coste*.

Coste	EER global(%)	DCF_{min}	EER medio (%)
0.01	25.92	0.0922	26.94
0.1	25.24	0.0923	26.25
1	24.18	0.0852	26
10	24.62	0.0834	26.2
100	25.54	0.0839	27.35
200	26.25	0.0842	27.72

Tabla 10: Resultados para varios *costes* para '*SUSAS Simulated* - AMF'.

Si se comparan los resultados de los sistemas *front-end* frente al sistema *back-end* de AMF [Tabla 11], lo primero que puede apreciarse es un incremento en el rendimiento de éste sistema frente a los primeros. Con AMF se consigue una EER_{media} de 26 % mientras que la fusión de los sistemas *front-end* [Tabla 11] obtiene un 30.46 %. Es decir, se reduce casi 4 puntos las tasas de error.

Se constata por lo tanto que nuestro nuevo sistema presentado en [28] logra mejorar los resultados del sistema GMM-SVM, SVM con estadísticos y la fusión suma.

En la Figura 33 se representa la curva DET para la fusión suma de los subsistemas *front-end* y la curva DET para el sistema de AMF.

Por último, la Tabla 11 analiza los EER_{medio} por emoción de tanto la fusión suma de los dos sistemas *front-end* como del sistema *back-end* de AMF. La última columna corresponde con la mejora relativa (M.R. en %) de éste último sistema con respecto al primero.

De la Tabla 11 es importante resaltar la gran diferencia de tasas de error entre emociones. Así, estilos de habla como *cond50 o cond70* tienen una tasa de error de reconocimiento muy

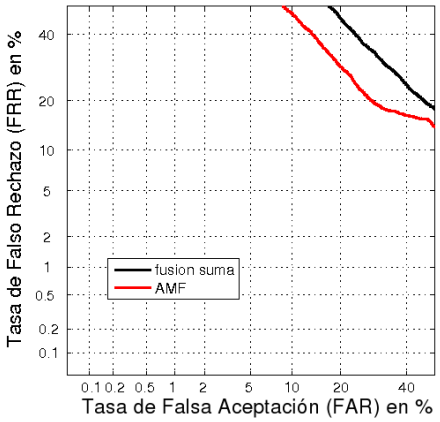

Figura 33: Curvas DET de la '*SUSAS Simulated* - fusión suma y AMF'.

Emoción	**EER (%) fusión suma**	**EER (%) AMF**	**M.R. (%)**
angry	19.93	21.88	+9.78
clear	34.85	34.84	-0.03
cond50	39.49	41.14	+4.18
cond70	45.30	33.41	-26.25
fast	33.84	22.10	-34.69
lombard	31.99	31.02	-3.03
loud	31.90	41.80	+31.03
neutral	40.43	10.05	-75.14
question	3.38	3.43	+1.48
slow	24.36	24.35	-0.04
soft	29.61	21.97	-25.8
EER_{medio}	30.46	26	-14.64

Tabla 11: EER (%) por emoción para '*SUSAS Simulated* - fusión suma y AMF'.

alta mientras que otros como *question* la tiene muy baja.

Una conclusión que se obtiene es que en emociones en las cuales hay una alta variación de la intensidad de habla, como *angry*, o una gran variación de la frecuencia fundamental, como *question*, se consiguen tasas de error relativamente bajas con respecto a la tasa media. Esto es debido a que justamente en nuestra parametrización hemos utilizado tanto la energía de habla como el pitch y sus correspondientes variaciones. Por lo tanto, si se quiere obtener mejores tasas de error en emociones en las que con la parametrización actual no se consiguen habría primero que analizar las propiedades prosódicas o acústicas que caracterizan a cada una de ellas y obtener un nuevo tipo de parametrización.

Otra conclusión que se puede sacar de la Tabla 11 es que aunque AMF mejora el rendimiento sobre casi todas las emociones, para *loud* (+31 %) y *angry* (+9.78 %), que son justamente los estilos con alta intensidad de habla, se produce un empeoramiento relativo con respecto a la fusión suma. La mayor mejora relativa ocurre en la emoción *neutral*, la cual pasa de un 40.43 % a un 10.05 %. Esto quiere decir que en el nuevo espacio de dimensiones de *Anchor Models* se consigue modelar mejor dicha emoción que en el espacio de parámetros inicial.

SUSAS Actual

Aquí vamos a ver los experimentos independientes de locutor realizados sobre la base de datos *SUSAS Actual*. Como ya sabemos, esta base de datos tiene 7 locutores los cuales se van a dividir también en 3 grupos según a que etapa (*development*, entrenamiento y test) se dediquen [Ver Tabla 12].

Etapa	**Locutores**
Development	*f1,m1*
Entrenamiento	*f2,m2*
Test	*f3,m3,m4*

Tabla 12: Distribución de locutores para experimentos independientes de locutor en *SUSAS Actual* .

- ***SUSAS Actual* - SVM con estadísticos**

Como se ve en la Figura 34, se entrenan 5 modelos ($\vec{w}_{SVM\_neutral}$, $\vec{w}_{SVM\_medst}$, $\vec{w}_{SVM\_hist}$, $\vec{w}_{SVM\_scream}$ y $\vec{w}_{SVM\_freefall}$), uno por cada emoción utilizando para ello los locutores de entrenamiento (*f2, m2*). El número de locuciones de entrenamiento por cada emoción es de:

35 palabras * 2 repeticiones/palabra * 2locutores = 140 locuciones/emoción.

Para este caso en que no se entrena un UBM no se usa los datos de los locutores *f1,m1*.

Una vez se tiene un modelo por cada emoción pasamos a la evaluación de los mismos. Para ello se usa los datos de test de los locutores *f3,m3,m4*. Se enfrentan todas las locuciones de test frente a los 5 modelos.

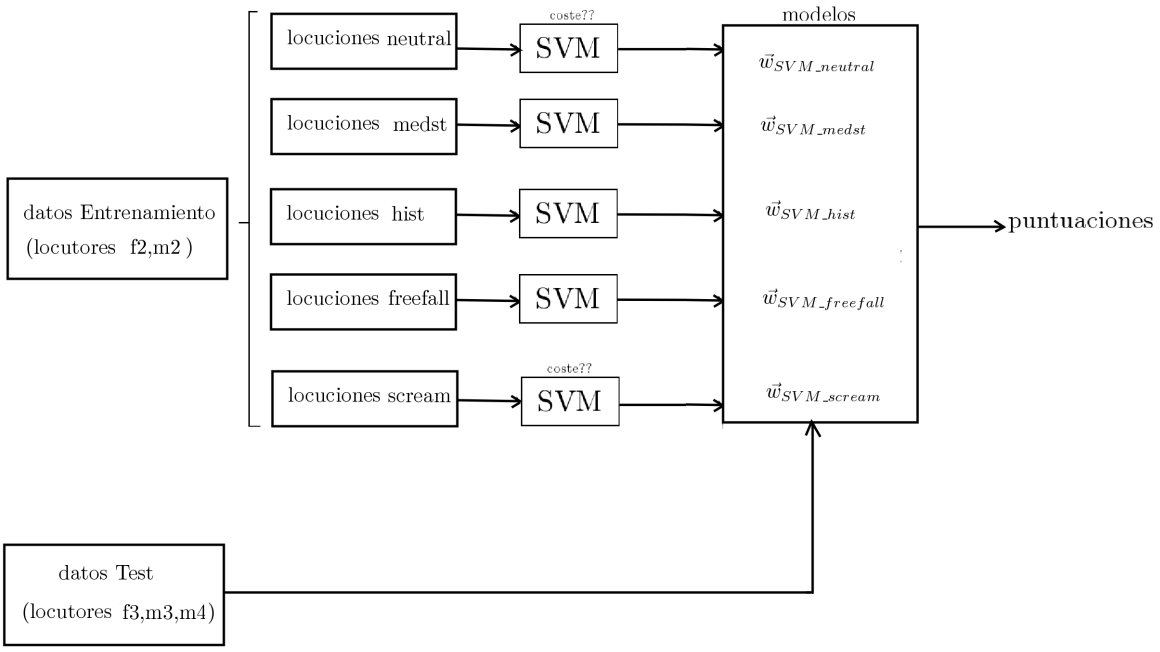

Figura 34: Esquema de las pruebas independientes de locutor para '*SUSAS Actual* - SVM con estadísticos'.

El número de locuciones de test es de:

35 palabras * 2 repeticiones/palabra * 3 locutores * 5 emociones = 1050 locuciones.

Por lo tanto, como cada locución de test se enfrenta a los 5 modelos, tendremos:

5 modelos * 1050 locuciones = 5250 puntuaciones.

Para este sistema se van a llevar a cabo las siguientes tareas para optimizar los resultados:

- Normalización de los vectores de parámetros prosódicos
- Optimización variable *coste* de entrenamiento
- T-normalización de puntuaciones

Para este tipo de experimentos y después de realizar varias pruebas con distintas normalizaciones de los parámetros prosódicos, se opta por no normalizar ningún vector de $\vec{u}_p$ pues es la opción que mejores resultados consigue. Por lo tanto se mantienen los vectores de parámetros originales $\vec{u}_p$:

$$\vec{u_p} = \{\vec{e}, \vec{p}, \vec{\Delta_e}, \vec{\Delta_p}\}$$

Como ya se hizo para *SUSAS Simulated*, se ajustará la variable *coste* del clasificador SVM.

Los resultados ya T-normalizados se muestran en la Figura 35 en forma de curvas DET y en la Tabla 13 con valores numéricos.

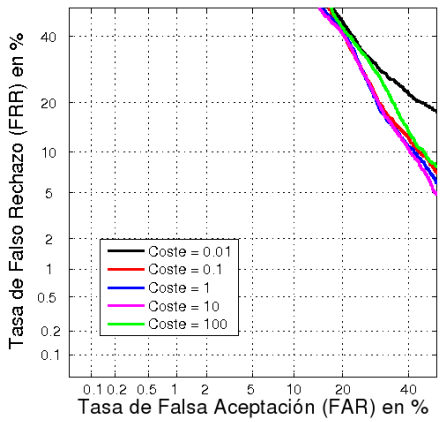

Figura 35: Curvas DET del sistema '*SUSAS Actual* - SVM con estadísticos' para diferentes *costes*.

Coste	**EER**$_{global}$	**DCF**$_{min}$	**EER**$_{medio}$
0.01	29.92	0.1	39.47
0.1	26.64	0.1	29.45
1	26.45	0.0996	28.89
10	26.54	0.0999	27.93
100	28.96	0.0998	27.93

Tabla 13: Resultados para '*SUSAS Actual* - SVM con estadísticos' dependiendo del *coste*.

Analizando la Tabla 13 se opta por un valor de *coste* 1 pues aunque con costes superiores se alcanza mejor EER$_{medio}$, ésto supone bastante mayor tiempo en entrenar los modelos SVM.

La configuración final para este tipo de pruebas se puede ver en la Tabla 14:

Normalización $\vec{u}_p$	**Coste**	**T-norm**	**EER**$_{global}$	**DCF**$_{min}$	**EER**$_{medio}$
no	1	sí	26.45	0.0996	28.89

Tabla 14: Configuración y resultados optimizados para '*SUSAS Actual* - SVM con estadísticos'.

- ***SUSAS Actual* - GMM-SVM**

Para el subsistema GMM-SVM, se usan los datos de *development* (f1, m1) para entrenar el modelo UBM que nos servirá como base para la posterior adaptación a los modelos GMM. El número de datos de *development* es de:

35 palabras * 2repeticiones/palabra * 2 locutores * 5emociones = 700 locuciones.

Los datos de entrenamiento (*f2,m2*) adaptaran dicho UBM generando así un modelo GMM por cada locución. [Ver Figura 36]

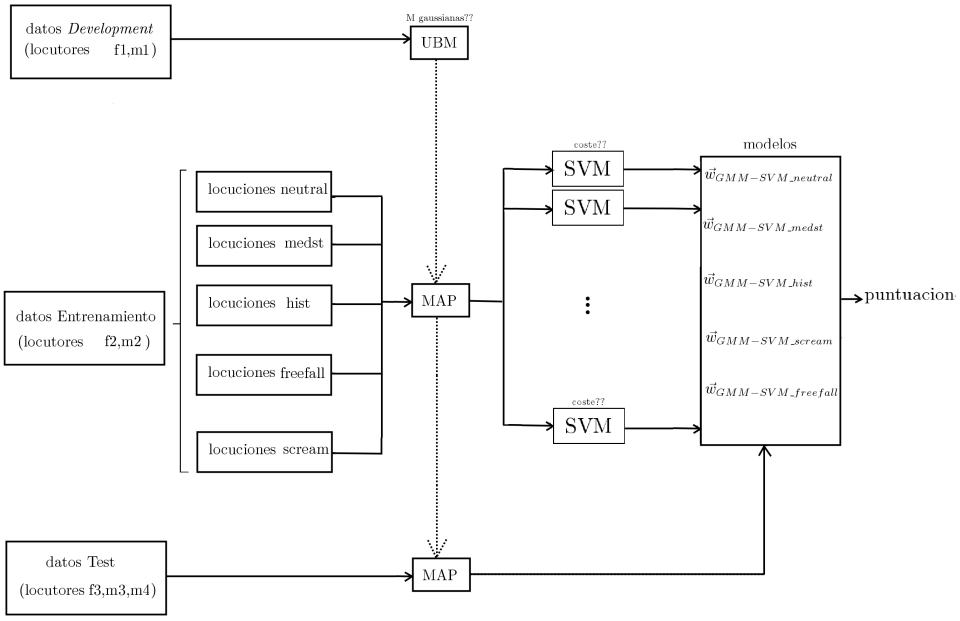

Figura 36: Esquema de las pruebas independientes de locutor para '*SUSAS Actual* - GMM-SVM'.

Como se ve en la Figura 36, mediante los clasificadores SVM se entrenan 5 modelos, uno por emoción ($\vec{w}_{GMM-SVM\_neutral}$, $\vec{w}_{GMM-SVM\_medst}$, $\vec{w}_{GMM-SVM\_hist}$, $\vec{w}_{GMM-SVM\_scream}$ y $\vec{w}_{GMM-SVM\_freefall}$). Al igual que el sistema de SVM con coeficientes estadísticos, se dispone de 140 locuciones de entrenamiento por emoción.

En este sistema las tareas que se van a realizar para optimizar resultados son:

- Normalización de los vectores de parámetros prosódicos
- Optimización variable M número de gaussianas
- Optimización variable *coste* de entrenamiento
- T-normalización de puntuaciones

Tras una serie de pruebas realizadas normalizando cada uno de los vectores de parámetros prosódicos de $\vec{u_p}$, la configuración que ofrece mejores resultados es mediante la normalización de tanto el vector de energías $\vec{e}$ como el de su velocidad $\vec{\Delta}_e$. Por lo tanto $\vec{u}_p$ queda:

$$\vec{u_p} = \{\vec{e} - E(\vec{e}), \vec{p}, \vec{\Delta_e} - E(\vec{\Delta_e}), \vec{\Delta_p}\}$$

Para optimizar la variable M, se han lanzado también una serie de pruebas para finalmente elegir un valor de M de 32 gaussianas. Contrasta que este valor sea mucho menor que las 256 gaussianas en este mismo tipo de experimentos para la base de datos *SUSAS Simulated*. Esto es debido a que al haber menos locutores para *SUSAS Actual*, la cantidad de datos es menor y por ello no se consigue modelar correctamente un número mezclas tan alto como son 256.

A continuación vamos a ajustar la variable coste manteniendo fijo el número de gaussianas a 32. La Figura 37 refleja los resultados para distintos valores de coste de entrenamiento, mientras que la Tabla 15 los muestra numéricamente.

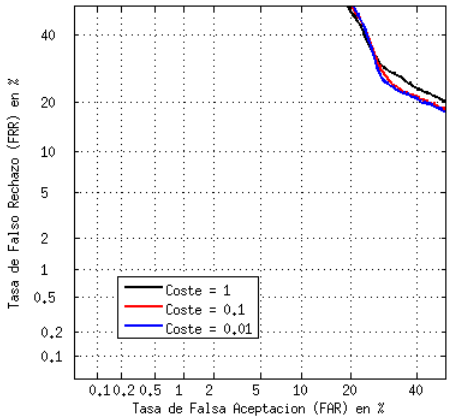

Figura 37: Curvas DET del sistema '*SUSAS Actual* - GMM-SVM' para diferentes *costes*.

Coste	**EER global(%)**	**DCF$_{min}$**	**EER medio (%)**
0.01	29.45	0.0998	36.5
0.1	29.82	0.0999	37.3
1	31.09	0.1	37.8

Tabla 15: Resultados del sistema '*SUSAS Actual* - GMM-SVM' dependiendo del *coste*.

Según los resultados de la Tabla 15, cuanto menor es el valor de la variable coste, mejores resultados se consiguen. Como ya se vio en el capítulo 3.7.2, un valor de coste muy pequeño hace priorizar la condición de maximizar el margen entre clases en el entrenamiento SVM. En la Tabla 16 aparece la configuración óptima para el subsistema GMM-SVM independiente de locutor con la base de datos *SUSAS Actual* .

Norm. $\vec{u}_p$	**M**	**coste**	***T-norm***	**EER$_{global}$**	**DCF$_{min}$**	**EER$_{medio}$**
$\vec{e}$ y $\vec{\Delta}_e$	32	0.01	sí	29.45	0.0998	36.5

Tabla 16: Configuración y resultados optimizados para '*SUSAS Actual* - GMM-SVM'.

- ***SUSAS Actual* - Fusión suma SVM estadísticos + GMM-SVM**

Para realizar la fusión suma de los resultados obtenidos mediante los dos subsistemas, GMM-SVM y SVM con estadísticos, se han utilizado las respectivas configuraciones de las Tablas 14 y 16 que ofrecen mejores resultados. Las curvas DET de tanto la fusión suma como

de los dos subsistemas aparecen en la Figura 38.

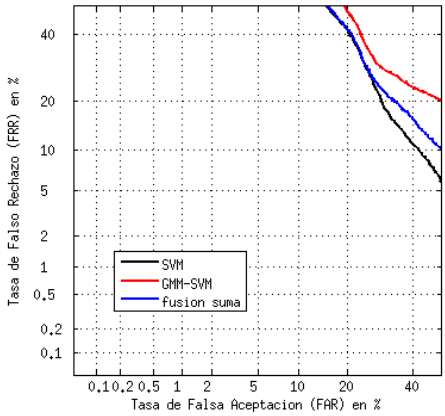

Figura 38: Curvas DET para '*SUSAS Actual* - SVM con estadísticos, GMM-SVM y fusión suma'.

Los valores de EER global de GMM-SVM, SVM con estadísticos y la fusión suma son 29.45 %, 26.45 % y 26.66 % respectivamente. Para este caso la fusión suma consigue resultados prácticamente iguales a los del mejor subsistema.

- ***SUSAS Actual* - Fusión de *Anchor Models* (AMF)**

Cada locución de test se enfrenta con los 5 modelos de cada uno de los subsistemas GMM-SVM y SVM con estadísticos para conformar un nuevo vector de parámetros de dimensión 10 ($\vec{S}_{locucion\_test}$) [Ver Figura 39].

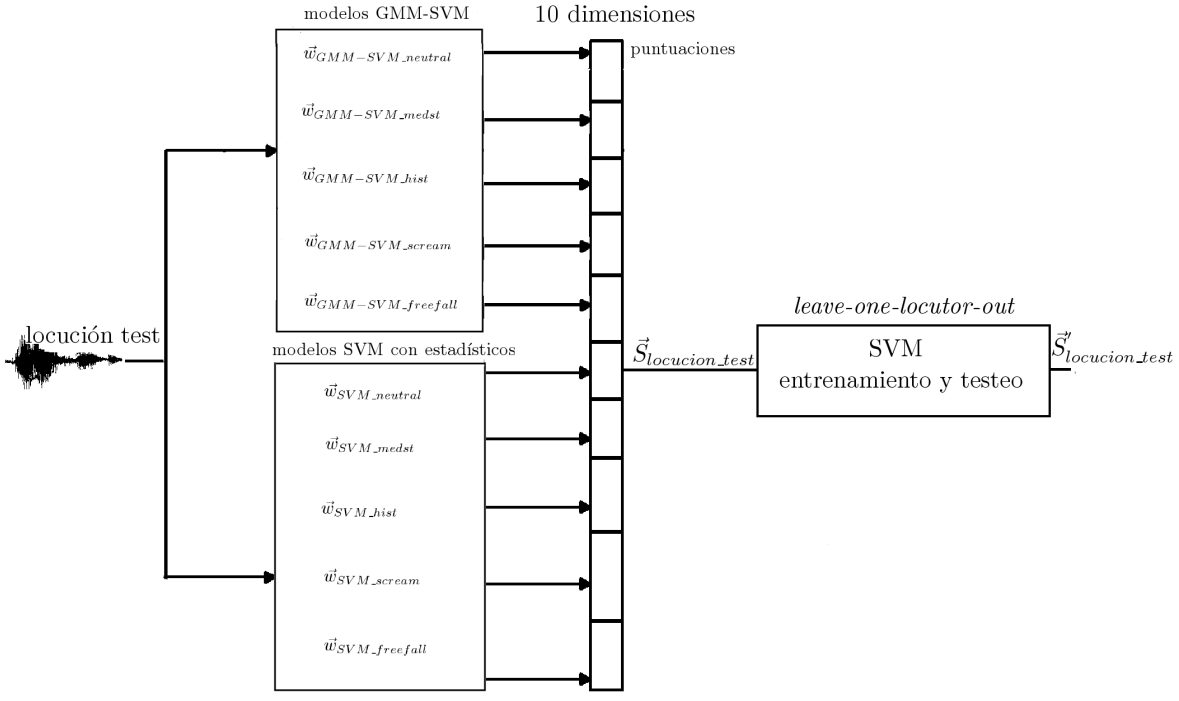

Figura 39: Esquema de las pruebas independientes de locutor para '*SUSAS Actual* - AMF'.

Los nuevos vectores de parámetros en el espacio de los *Anchor Models* $\vec{S}_x$ se utilizan como entrada a un clasificador SVM. Igual se hizo para *SUSAS Simulated*, se aplica la técnica de

leave-one-locutor-out. Es decir, se seleccionarán iterativamente los datos de cada uno de los tres locutores *f3,m3,m4* y se utilizarán para evaluación mientras que los datos de los otros dos restantes se utilizarán para entrenar un modelo SVM por emoción.

En la Figura 40 se representan un conjunto de curvas DET para varios valores de la variable *coste* del clasificador *back-end* SVM. Y en la Tabla 17 valores numéricos de tasas de error y DCF_{min}. Como se puede ver en dicha tabla, el valor de coste que optimiza los resultados es de 1 si se opta por el mejor valor de EER global o 0.1 si se desea el valor óptimo de EER_{medio}.
.

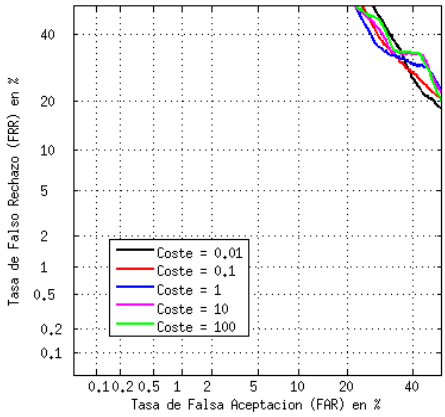

Figura 40: Curvas DET para '*SUSAS Actual* - AMF' para varios valores de *coste*.

Coste	**EER global(%)**	**DCF_{min}**	**EER medio (%)**
0.01	35.41	0.099	33.11
0.1	33.37	0.0987	32.46
1	33.03	0.0988	35.70
10	34.05	0.0991	37.20
100	34.39	0.0995	37.80

Tabla 17: Resultados dependiendo del *coste* '*SUSAS Actual* - AMF'.

En *SUSAS Actual* , a diferencia de lo que ocurría en *SUSAS Simulated*, el sistema *back-end* de AMF empeora con respecto a los subsistemas. Con AMF se consigue una EER_{media} de 35.7 % mientras que la fusión de los subsistemas *front-end* [Tabla 18] obtenía un 29.9 %. Es decir, AMF empeora en casi 6 puntos la EER_{media}. En la Figura 41 se representan la curva DET para la fusión suma del sistema SVM de estadísticos con el sistema GMM-SVM y la curva DET para el sistema de AMF.

Por último, la Tabla 18 analiza los EER_{medio} por emoción de tanto la fusión suma de los dos subsistemas *front-end* como del sistema *back-end* de AMF. Al igual que en la Tabla 33, la cuarta columna muestra la mejora relativa (M.R. en %) que ofrece AMF frente a la fusión suma *front-end*

Una conclusión que se obtiene de la Tabla 18 es que a diferencia de lo que ocurría para la base de datos *SUSAS Simulated*, AMF no hace mejorar los resultados alcanzados por la fusión suma para *SUSAS Actual*.

Al igual que en *SUSAS Simulated*, hay estilos de habla o emociones que obtienen mejores tasas de error. Este es el caso del estilo *scream*. Las locuciones de dicho estilo de habla se caracterizan por tener una alta intensidad de voz (o energía) y gran variabilidad de la misma.

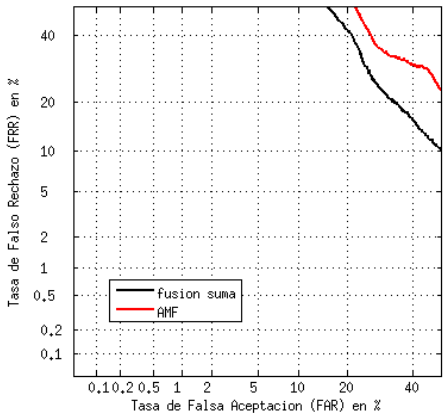

Figura 41: Curvas DET para '*SUSAS Actual* - fusión suma y AMF'.

Emoción	**EER (%) fusión suma**	**EER (%) AMF**	**M.R. (%)**
neutral	33.23	40.42	+21,64
medst	41.51	43.95	+5,88
hist	35.83	44.83	+25,12
freefall	31.29	37.57	+20,07
scream	7.64	11.72	+53,4
EER_{medio}	29.9	35.7	+19,4

Tabla 18: EER (%) por emoción para '*SUSAS Actual* - fusión suma y AMF'.

Es decir, los vectores prosódicos $\vec{e}$ y $\vec{\Delta}_e$ van a caracterizar bien dicha clase.

Ah3R1

En este apartado se van a describir los experimentos independientes de locutor realizados sobre la base de datos *Ah3R1*. Como se vio en el capítulo 4.1.1, esta base de datos dispone de 69 locutores. Cada uno ellos tiene un conjunto de locuciones para entrenamiento/*development* y otro para evaluación.

- ***Ah3R1* - SVM con estadísticos**

Para las dos bases de datos de SUSAS se dividían los locutores según la tarea a la que se emplearan sus locuciones.

Por el contrario, para *Ah3R1* los 69 locutores se van a emplear tanto para tareas de *development*/entrenamiento como de test. Así, la manera de generar los modelos SVM es la siguiente y es la que aparece en la Figura 42. Se van a entrenar modelos de la forma $\vec{w}_{SVM\_notLocX\_emoc}$. Dichos modelos serán entrenados con datos de la emoción *emoc* (*neutro-bajo, neutro, neutro-exaltado, exaltado*) utilizando locuciones de entrenamiento de todos los locutores menos el locutor X. Por lo tanto el número de modelos que serán entrenados es de:

4 emociones * 69 locutores = 276 modelos.

Una vez se ha generado un modelo por cada emoción se pasa a la etapa de evaluación de los mismos. Para ello se usa los datos de test de cada locutor. El procedimiento es el siguiente. Se evalúan las locuciones de test del locutor X frente a los modelos de la forma $\vec{w}_{SVM\_notLocX\_emoc}$, donde *emoc* es cada una de las 4 emociones de *Ah3R1*. De esta manera se realizan pruebas

independientes de locutor donde los datos de test de un locutor no se usan para evaluar modelos entrenados por ese mismo locutor.

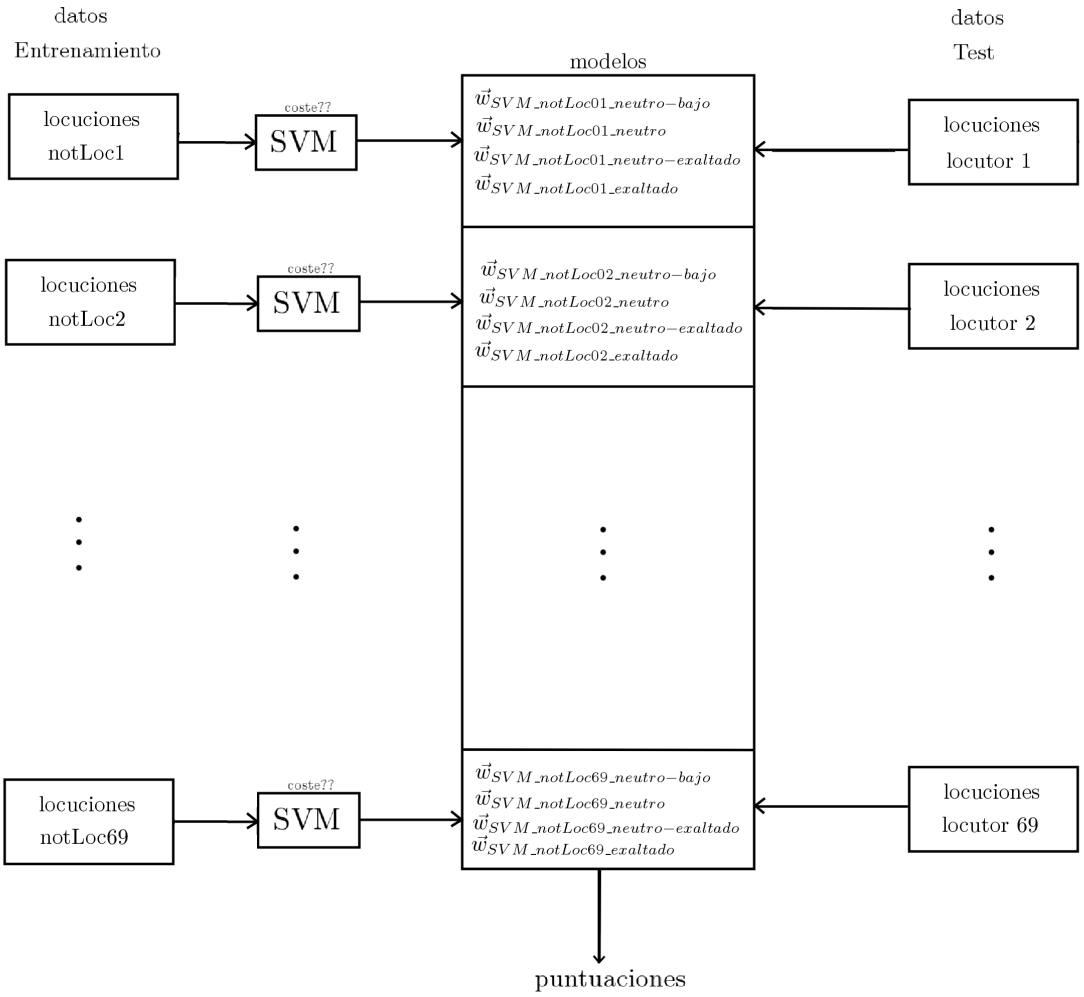

Figura 42: Esquema de las pruebas independientes de locutor para '*Ah3R1* - SVM con estadísticos'.

Como ya vimos en 4.1.1, no todos los locutores tienen el mismo número de locuciones de test. Así, los 31 primeros tienen 10 y los 38 restantes únicamente 5. Por lo tanto, el número de locuciones de test es de:

31 locutores * 10 locuciones/locutor + 38 locutores * 5 locuciones/locutor = 500 locuciones.

Como cada locución de test se enfrenta a cada uno de los 276 modelos, el número de puntuaciones será de:

276 modelos * 500 locuciones = 138000 puntuaciones.

Para este sistema se van a llevar a cabo las siguientes tareas para optimizar los resultados:

- Optimización variable coste de entrenamiento
- Normalización de los vectores de parámetros prosódicos
- T-normalización de puntuaciones

El primer valor a ajustar es la variable *coste* del entrenamiento SVM. Para ello mantenemos los vectores de parámetros prosódicos sin ningún tipo de normalización. Los resultados para varios valores de *coste* se representan en la Figura 43 y en la Tabla 19

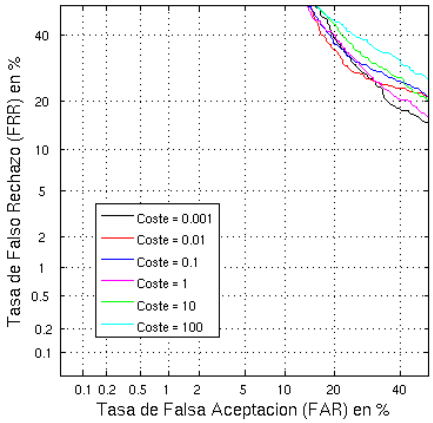

Figura 43: Curvas DET del sistema '*Ah3R1* - SVM con estadísticos' para diferentes *costes*.

Coste	EER global(%)	DCF_{min}	EER medio (%)
0.001	27.64	0.1	46.97
0.01	27.23	0.0997	44.56
0.1	29.28	0.0993	43.05
1	28.25	0.0997	37.32
10	30.51	0.0997	36.97
100	33.99	0.0997	37.79

Tabla 19: Resultados dependiendo del valor del *coste* para '*Ah3R1* - SVM con estadísticos'.

A la vista de los resultados de la Tabla 19, el valor del *coste* que hace minimizar el EER global no coincide con el que minimiza el EER medio. Ambas tasas de error únicamente coinciden cuando el número de enfrentamientos frente a cada modelo es el mismo. El que en *Ah3R1* ambas tasas difieran tanto es debido a la descompensación del número de locuciones según para que emoción. Así, existen muchas más locuciones de test de la emoción *neutro* que de *neutro-bajo o exaltado*.

Para un valor de *coste* muy pequeño como es 0.01 se consigue la mejor tasa de EER global de 27.23 % pero sin embargo el EER medio aumenta hasta el 44.46 %. Por otro lado, si se toma como *coste* el valor 10 se obtiene el mínimo EER medio de 36.97 % pero el EER global alcanza el 30.51 %. Por lo tanto, nos vamos a decantar por una opción intermedia como es *coste* 1, pues únicamente es 1 punto más alto que el mejor EER global logrando también uno de los mejores EER medio.

Una vez se ha ajustado el valor del *coste* a 1, lo siguiente es la normalización de los vectores prosódicos. En la Tabla 20 aparecen los resultados de varios experimentos según el vector o vectores prosódicos normalizados, manteniendo el valor de *coste* fijo a 1.

Según los resultados de la Tabla 20, se opta por elegir la opción de normalizar tanto el vector de energías $\vec{e}$ como el de su velocidad $\vec{\Delta}_e$ pues consigue reducir tanto la EER global como la media.

Tras los ajustes anteriores y la posterior T-normalización de las puntuaciones se obitienen los resultados de la Tabla 21:

- ***Ah3R1* - GMM-SVM**

Para la técnica de GMM-SVM, se usan todos los datos de entrenamiento para entrenar el

Normalización $\vec{u}_p$	EER global(%)	DCF$_{min}$	EER medio (%)
no	28.25	0.0997	37.32
$\vec{e}$	28.25	0.0997	34.16
$\vec{e}$ y $\vec{\Delta}_e$	27.78	0.1	34.24
$\vec{p}$	31.60	0.0997	40.76
$\vec{p}$ y $\vec{\Delta}_p$	31.87	0.0997	40.41
$\vec{e}$, $\vec{\Delta}_e$, $\vec{p}$ y $\vec{\Delta}_p$	30.71	0.0997	37.22
$\vec{e}$, $\vec{\Delta}_e$, y $\vec{\Delta}_p$	28.25	0.1	34.55

Tabla 20: Resultados para '*Ah3R1* - SVM con estadísticos' dependiendo de los vectores de parámetros prosódicos normalizados.

Norm. $\vec{u}_p$	Coste	T-norm	EER$_{global}$	DCF$_{min}$	EER$_{medio}$
$\vec{e}$ y $\vec{\Delta}_e$ y	1	sí	27.44	0.0991	32.95

Tabla 21: Configuración y resultados optimizados para '*Ah3R1* - SVM con estadísticos'.

modelo UBM. Cada locución de entreamiento lo adaptará para así generar un modelo GMM por cada locución. [Ver Figura 44]

La manera de entrenar los modelos [Ver Figura 42] es la misma que para el caso anterior de SVM con estadísticos. Se entrenan modelos de la forma $\vec{w}_{GMM-SVM\_notLocX\_emoc}$. Dichos modelos serán entrenados con datos de la emoción *emoc* (*neutro-bajo, neutro, neutro-exaltado, exaltado*) utilizando locuciones de entrenamiento de todos los locutores menos el locutor X.

En la etapa de evaluación se testean las locuciones de test del locutor X frente a los modelos de la forma $\vec{w}_{GMM-SVM\_notLocX\_emoc}$, donde *emoc* es cada una de las 4 emociones de *Ah3R1*.

Se va a optimizar sobre los siguientes parámetros:

- Normalización de los vectores de parámetros prosódicos
- Optimización variable M número de gaussianas
- Optimización variable coste de entrenamiento
- T-normalización de puntuaciones

El primer ajuste que se realizará es el del número de gaussianas M. Para ello se manteniene fijo el valor del *coste* a 0.1 y vamos variando M con valores potencia de 2.

Los resultados de esta optimización se muestran en la Tabla 22.

M	EER global(%)	DCF$_{min}$	EER medio (%)
64	24.03	0.0963	36.74
128	23.89	0.0938	32.65
256	23.95	0.0943	35.28
512	24.98	0.0935	33.96

Tabla 22: Resultados para '*Ah3R1* - GMM-SVM' variando el número de gaussinas.

A la vista de los resultados, resulta evidente que el número de gaussianas que hace que se obtengan mejores resultados es de 128, valor para el cual se minimizan tanto el EER global, como el DCF$_{min}$ como el EER medio.

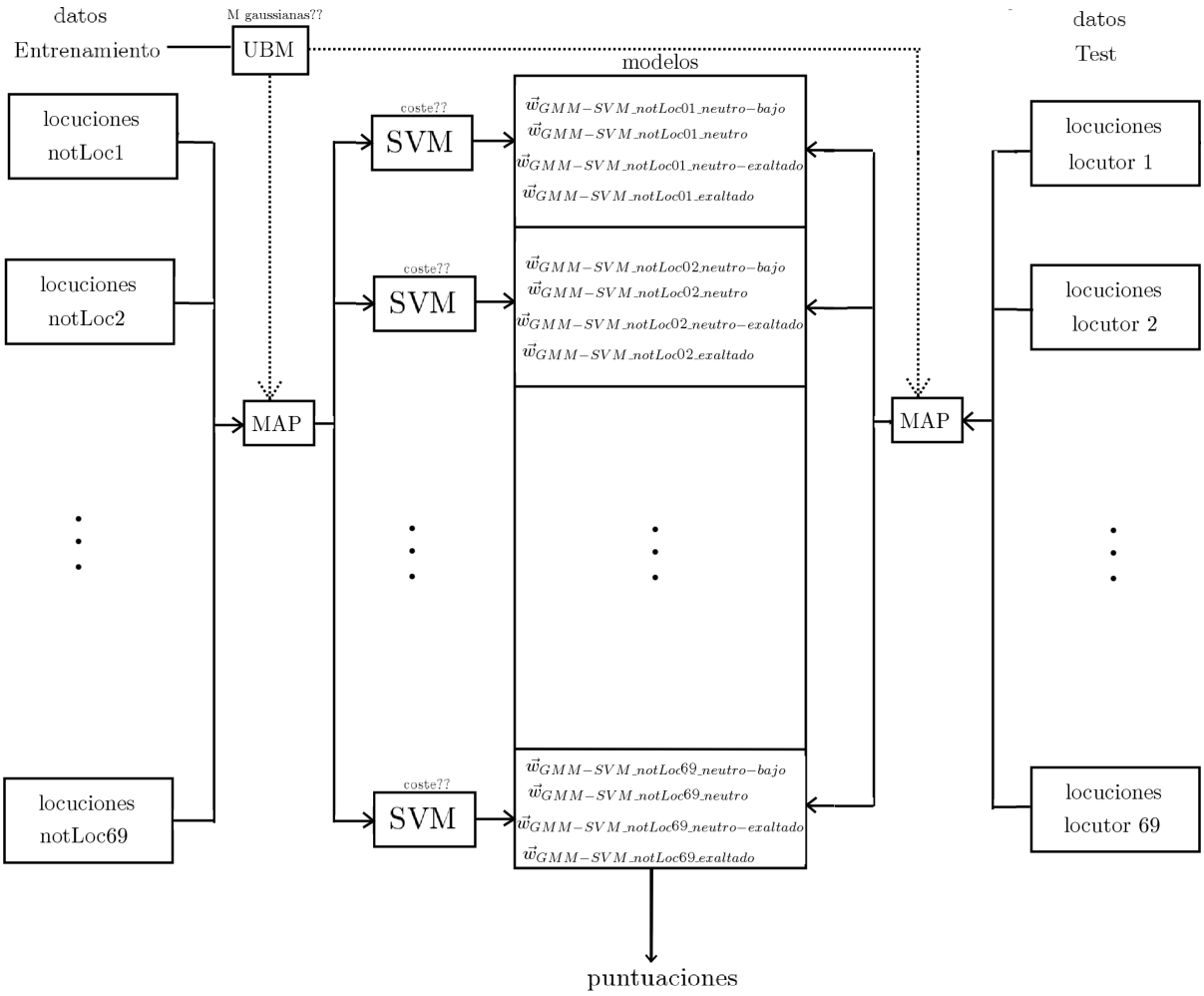

Figura 44: Esquema de las pruebas independientes de locutor para '*Ah3R1* - GMM-SVM'.

La siguiente variable a ajustar es el *coste*. Para ello se mantiene fijo el valor de M a 128 y se va variando el *coste*. La Figura 45 y la Tabla 23 muestran dichos resultados.

Coste	**EER global(%)**	**DCF$_{min}$**	**EER medio (%)**
0.001	27.84	0.0945	38.76
0.01	23.89	0.0948	40.44
0.1	23.89	0.0938	32.65
1	25.32	0.0961	34.10
10	27.78	0.0997	35.24

Tabla 23: Resultados dependiendo del *coste* para '*Ah3R1* - GMM-SVM'.

El valor de *coste* óptimo es de 0.1 pues minimiza tanto el EER medio, como el EER global, como el DCF$_{min}$.

Por último, los resultados terminan de de ser ajustados mediante la normalización o no de cada uno de los 4 vectores de parámetros prosódicos $\vec{e}$, $\vec{\Delta}_e$, $\vec{p}$ y $\vec{\Delta}_p$. Los resultados, para un valor fijo de *coste* y M de 0.1 y 128 respectivamente, de dichas normalizaciones aparecen en la Figura 46 y en la Tabla 24.

A partir de la Tabla 24, la configuración que logra optimizar los resultados es mediante la normalización de los vectores de energía $\vec{e}$ y su velocidad $\vec{\Delta}_e$.

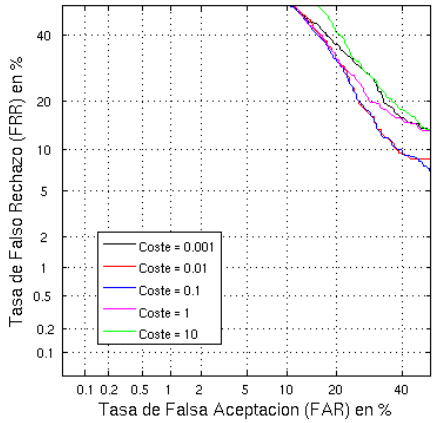

Figura 45: Curvas DET para varios *costes* para '*Ah3R1* - GMM-SVM'.

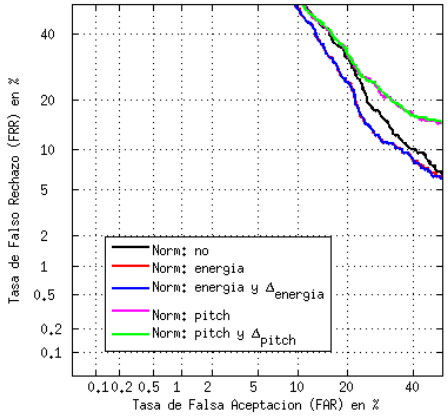

Figura 46: Curvas DET para '*Ah3R1* - GMM-SVM' según la normalización de los vectores de parámetros prosódicos.

Tras haber optimizado tanto el valor de coste, como el de M, como la normalización de los vectores prosódicos, la última tarea es la T-normalización de los resultados utilizando la mejor configuración [Ver Tabla 25].

Aquí, a diferencia de lo que ocurría en los casos anteriores, las tasas de error empeoran cuando se lleva a cabo la T-normalización de puntuaciones.

- **Ahumada III - Fusión suma SVM estadísticos + GMM-SVM**

Normalización $\vec{u}_p$	**EER global(%)**	**DCF$_{min}$**	**EER medio (%)**
no	23.89	0.0938	32.65
$\vec{e}$	21.63	0.0943	30.99
$\vec{e}$ y $\vec{\Delta}_e$	21.63	0.0943	30.88
$\vec{p}$	25.59	0.0993	43.54
$\vec{p}$ y $\vec{\Delta}_p$	25.18	0.0993	41.92

Tabla 24: Resultados dependiendo de los vectores de parámetros prosódicos normalizados para '*Ah3R1* - GMM-SVM'.

Norm. $\vec{u}_p$	**M**	**coste**	***T-norm***	**EER**$_{global}$	**DCF**$_{min}$	**EER**$_{medio}$
$\vec{e}$ y $\vec{\Delta}_e$	128	0.1	sí	25.52	0.0933	33.92

Tabla 25: Configuración y resultados optimizados para '*Ah3R1* - GMM-SVM'.

Se toman los resultados de las Tablas 21 y 25 como los resultados óptimos para los subsistemas *front-end* de SVM con estadísticos y GMM-SVM respectivamente. Dichos resultados están T-normalizados para que los rangos de puntuaciones de ambos subsistemas sean parecidos.

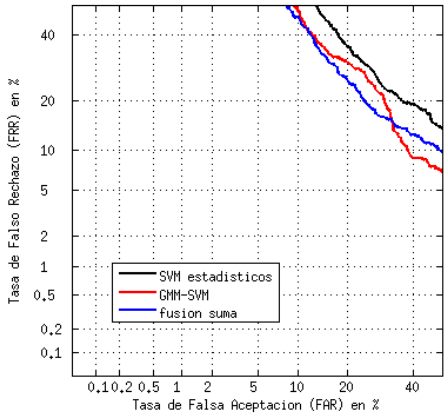

Figura 47: Curvas DET de '*Ah3R1* - SVM con estadísticos, GMM-SVM y fusión suma'.

Las tasas de EER global de tanto el subsistema SVM con estadísticos como el de GMM-SVM como la fusión suma de ambos son de 27.44 %, 25.52 % y 22.59 % respectivamente. Como se ven claramente en la Figura 47, la fusión suma consigue reducir notablemente las tasas de error de los subsistemas.

- ***Ah3R1* - Fusión de *Anchor Models* (AMF)**

Cada locución de test se enfrenta con los 4 modelos de cada uno de los subsistemas GMM-SVM y SVM con estadísticos que no han sido entrenados con datos de ese mismo locutor. Así, se forma un nuevo vector de parámetros de dimensión 8 ($\vec{S}_{locucion\_test}$) [Ver Figura 48].

El nuevo vector de puntuaciones de dimensión 8 corresponde con nuestro nuevo vector de parámetros. Dicho vector será nuestro supervector que servirá para modelar un nuevo clasificador SVM. En *Ah3R1*, al igual que hicimos en AMF para la base de datos SUSAS, se aplicará la validación cruzada *leave-one-locutor-out*. Se cogerán iterativamente los datos de cada uno de los 69 locutores *Loc01,...,Loc69* y se utilizarán para evaluación mientras que los datos de los 68 restantes se utilizarán para entrenar un modelo SVM por emoción.

En la Figura 49 se representan curvas DET para varios valores de la variable coste del clasificador *back-end* SVM. Y en la Tabla 26 valores numéricos de tasas de error y DCF$_{min}$.

.

Teniendo en cuenta el EER global podríamos decir que el valor de *coste* óptimo es de 0.1 pues alcanza un 21.17 % cosa que otro valor de *coste* no lo alcanza. Sin embargo, se aprecia que con un *coste* de 10 apenas empeora el EER global mejorando 3 puntos el EER medio. Por esa razón elegimos dicho valor de *coste* aunque el tiempo de entrenamiento sea mayor.

A tenor de los resultados anteriores se puede decir que para la base de datos *Ah3R1* la técnica *back-end* de AMF apenas consigue mejorar los resultados que ofrece la fusión de los dos subsistemas *front-end*. En concreto la fusión suma obtiene un EER global de 22.59 %

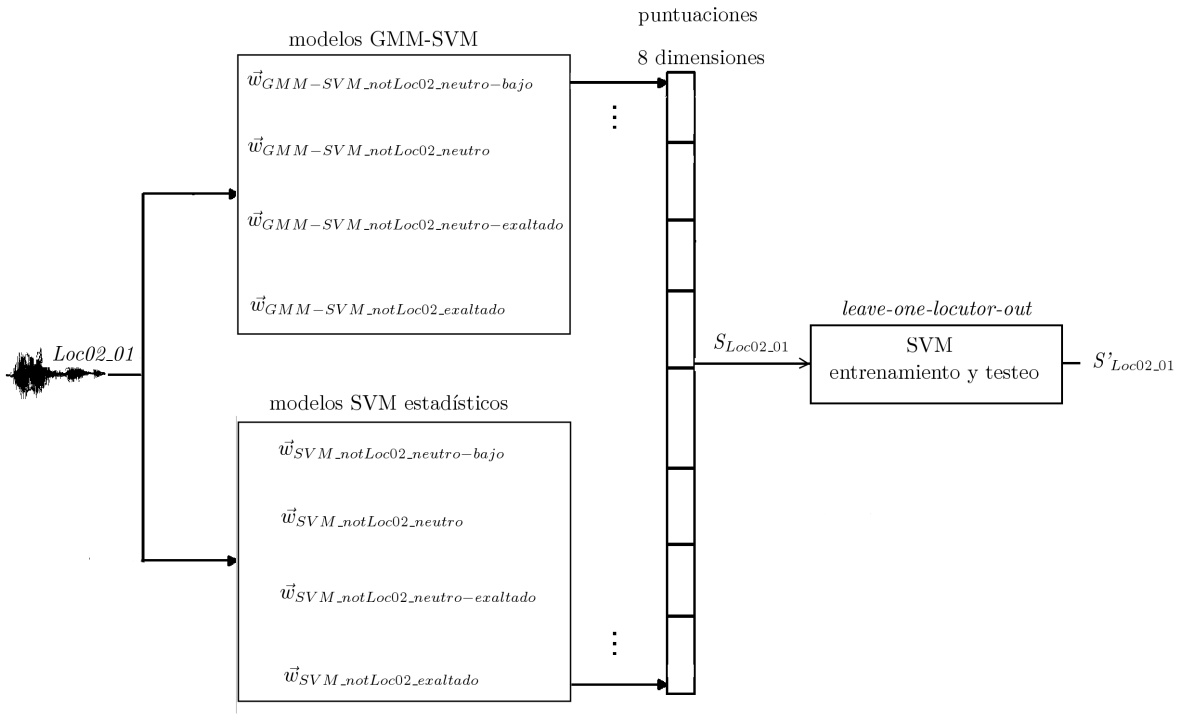

Figura 48: Esquema de las pruebas independientes de locutor para '*Ah3R1* - AMF'.

Coste	**EER global(%)**	**DCF$_{min}$**	**EER medio (%)**
0.01	22.21	0.0995	36.48
0.1	21.17	0.0985	34.62
1	22.83	0.099	35.57
10	22.83	0.0987	31.65
100	23.04	0.0994	32.45

Tabla 26: Resultados dependiendo del *coste* para '*Ah3R1* - AMF'.

por un 22.83 % de AMF. Si se mide en EER medio se pasa de un 34.01 % a un 31.65 % con AMF.

En la Figura 50 se representa la curva DET de la fusión suma y de AMF. Mientras que en la Tabla 27 se analizan los EER medios por emoción para ambas técnicas. También se muestran las mejoras relativas (M.R. en %) que ofrece AMF frente a la fusión suma.

Según la Tabla 27, AMF ofrece una mejora en el EER medio que no llega a los 3 puntos con respecto a la fusión suma. Esto supone una mejora relativa del -6.94 % puntos. La mejora relativa que se conseguía para las bases de datos *SUSAS Simulated* y *SUSAS Actual* era del -14.64 % y +19,4 %. Es decir, mientras que para tanto *SUSAS Simulated* como *Ah3R1* la técnica de AMF mejora con respecto a los sistemas *front-end*, en *SUSAS Actual* empeora considerablemente.

Entre los resultados de *Ah3R1* y los de SUSAS se aprecia una diferencia. Para los primeros no hay tanta diferencia entre emociones mientras que en SUSAS había emociones como *angry, question o scream* con las que se obtenían mucho mejores tasas de error que para el resto. Posiblemente esto es debio a que en *Ah3R1* las emociones o estilos de habla (*neutro-bajo,*

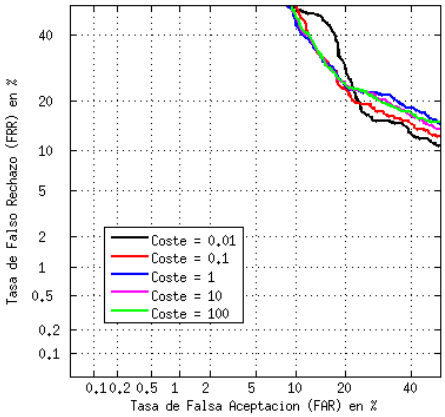

Figura 49: Curvas DET del sistema '*Ah3R1* - AMF' según la variable *coste*.

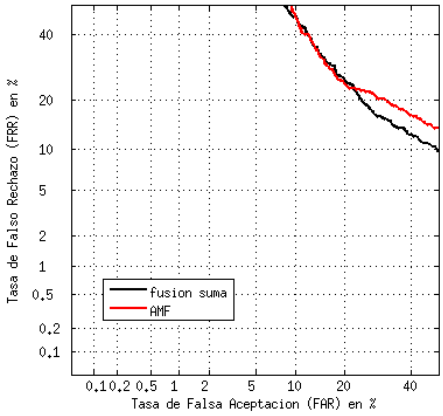

Figura 50: Curvas DET para '*Ah3R1* - fusión suma y AMF'.

neutro, neutro-exaltado y exaltado) están mucho menos definidas o cubren un rango más amplio que las de las bases de datos de SUSAS. Aún con eso, en *Ah3R1* se pueden apreciar ligeras diferencias de tasas errores según la emoción. Así, la emoción *exaltado* es la que mejores resultados ofrece llegando a un EER del 25.05 % para la fusión suma o el 31.65 % para AMF. Lo que es común para las 3 bases de datos es que los estilos de habla que se caracterizan por una alta intensidad de habla o de frecuencia (*angry o question* en *SUSAS Simulated*, *scream o freefall* en *SUSAS Actual* y *exaltado* en *Ah3R1*) funcionan mucho mejor que el resto. Por algo nuestros vectores paramétricos incluyen la energía y la frecuencia fundamental.

Una vez se han visto los resultados para experimentos independientes de locutor, se van

Emoción	**EER (%) fusión suma**	**EER (%) AMF**	**M.R. (%)**
neutro-bajo	38.87	27.86	-28.33
neutro	34.82	33.48	-3.85
neutro-exaltado	37.30	34.28	-8.1
exaltado	25.05	30.97	+23.63
EER_{medio}	34.01	31.65	-6.94

Tabla 27: EER (%) por emoción para '*Ah3R1* - fusión suma y AMF'.

a comparar según la base de datos. Así, en la Tabla 28 nos muestra el EER medio para las 3 bases de datos.

base de datos	***front-end/back-end***	**EER medio (%)**
SUSAS Simulated	front-end(fusión suma)	30.46
	back-end (AMF)	26.00
SUSAS Actual	front-end(fusión suma)	29.90
	back-end (AMF)	35.70
Ah3R1	front-end(fusión suma)	34.01
	back-end (AMF)	31.65

Tabla 28: EER_{medio} (%) para las 3 bases de datos para experimentos independientes de locutor.

Viendo la Tabla 28 de resumen, la base de datos en que la técnica de AMF consigue mejoras considerables con respecto a la fusión suma es *SUSAS Simulated*. Posiblemente eso sea debido a que dicha base de datos está formada por locuciones de habla de emociones simuladas, posiblemente exageradas. Así el espacio de *Anchor Models* en el que trabaja AMF es mucho más discriminativo para esta base de datos.

Por otro lado, la base de datos sobre la que se han obtenido mejores resultados, tanto de AMF como de la fusión suma, es también *SUSAS Simulated*. La razón es la misma, aunque es la que más emociones tiene, las emociones están exageradas y claramente diferenciadas unas de otras. Así, se puede concluir que nuestros sistemas para tareas independientes de locutor discriminan mejor sobre un conjunto amplio de clases o emociones bien diferenciadas o exageradas que sobre un conjunto más pequeño pero más confusas.

5.1.2. Experimentos *Inter*-Base de datos: Evaluación de cada Base de Datos frente a modelos de todas las Bases de Datos

Este capítulo trata de, en vez de evaluar cada base de datos con modelos creados con datos de la misma base de datos, evaluar cada una con modelos de todas las bases de datos. Así, por ejemplo, las locuciones de test de *SUSAS Simulated* se enfrentará con modelos de tanto *SUSAS Simulated*, como de *SUSAS Actual* como de *Ah3R1*.

Para las bases de datos *SUSAS Simulated* y *SUSAS Actual* se han entrenado un modelo por cada emoción. Son 11 (*angry, clear, cond50, cond70, fast, lombard, loud, neutral, question, slow y soft*) para la bases de datos *SUSAS Simulated* y 5 (*neutral, medst, hist, freefall y scream*) para la bases de datos *SUSAS Actual*. Sin embargo, para la base de datos *Ah3R1* al haber hecho *cross validation* no tenemos un modelo por cada emoción, sino un modelo por cada emoción y locutor. Por lo tanto existen 276 modelos (276 = 4 emociones * 69 locutores).

Se podría tomar los 276 modelos de *Ah3R1* pero se hiciese habría una gran descompensación entre el número de modelos por cada base de datos. Por lo tanto se toman 4 modelos cualquiera de los 276 de *Ah3R1*. Uno de cada emoción (*neutro-bajo, neutro, neutro-exaltado y exaltado*).

Entre las tres bases de datos se dispone por lo tanto de 20 modelos por cada subsistema *front-end* (11 de *SUSAS Simulated*, 5 de *SUSAS Actual* y 4 de *Ah3R1*). La Figura 51 muestra la forma en que se va a evaluar cada uno de estos modelos.

La parte de datos de cada base de datos reservada para test se usa para evaluar dichos modelos. Para *SUSAS Simulated* se reservan los locutores *g3, b3 y n3* para dicha tarea. Para *SUSAS Actual* los locutores *f3, m3 y m4*. Y para *Ah3R1* se usan las locuciones de test que

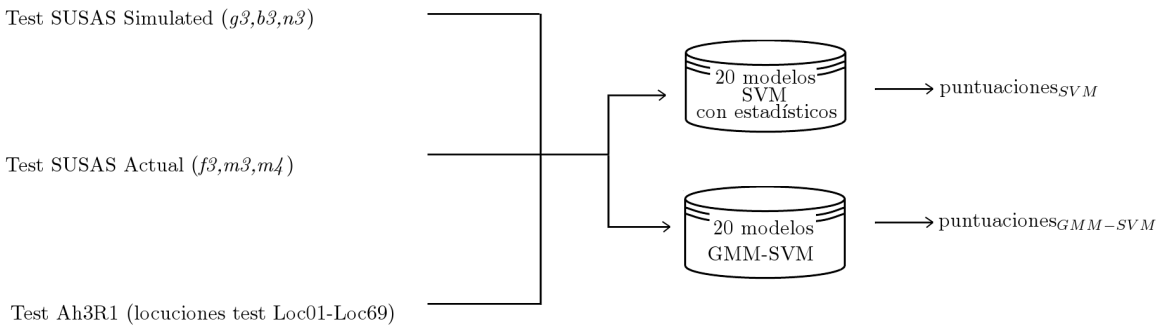

Figura 51: Esquema de evaluación de los modelos de las 3 bases de datos.

hay para cada uno de los 69 locutores. En concreto, existen 10 locuciones de test para los 31 primeros locutores y 5 para los 38 restantes.

Una vez se tiene claro el número de modelos que se van a tomar por cada base de datos, lo siguiente es decidir la configuración a establecer para entrenar dichos modelos. Es decir, valores de *coste*, M, tipo de normalización de los vectores prosódicos, etc. A priori se podrían tomar aquellas configuraciones que han dado mejores resultados. Sin embargo, si se hiciese eso, los modelos de distintas bases y datos y subsistemas tendrían distinta configuración y por lo tanto habría incompetencia entre modelos. Por ello, decidimos por entrenar todos los modelos con la siguiente configuración:

· Normalización vectores parámetros prosódicos: no
· M, número de Gaussianas: 256
· *coste*: 1.

Una vez se ha sacado los resultados T-normalizados para los dos subsistemas *front-end* se hace la fusión suma.

Por último se realiza AMF. Cada locución de test de las 3 bases de datos se enfrenta con los 20 modelos de cada uno de los subsistemas GMM-SVM y SVM con estadísticos. Así, se forma un nuevo vector de parámetros de dimensión 40.

El nuevo vector de puntuaciones de dimensión 40 será nuestro nuevo vector de parámetros. Dicho vector será nuestro supervector que servirá para modelar un nuevo clasificador SVM al cual se le ajustará el *coste*. Se cogerán iterativamente los datos de cada locutor y se utilizarán para evaluación mientras que los datos de los restantes locutores se utilizarán para entrenar los modelos SVM.

La Tabla 29 nos ofrece los resultados de tanto los subsistemas *front-end*, como de la fusión suma de ambos, como de AMF.

Como se dijo anteriormente, se ha ajustado la variable *coste* para el clasificador SVM del sistema *back-end*. Tras realizar los experimentos se ha visto que para un valor de *coste* 1 se optimizan los resultados.

Se puede ver en la Tabla 29 que para estos tipos de experimentos *Inter*-Base de datos, la técnica de AMF consigue mejorar en todos los casos los resultados de la fusión suma. Este hecho refleja que los AMF funcionan mejor cuanto mayor dimensión del espacio *Anchor Model*

	SVM estadísticos		**GMM-SVM**		**fusión suma**		**AMF**	
	EER_{glob}	EER_{med}	EER_{glob}	EER_{med}	EER_{glob}	EER_{med}	EER_{glob}	EER_{med}
Simulat.	39.39	34.78	36.79	29.44	39.15	31.01	27.74	28.80
Actual	29.62	32.46	37.06	51.10	25.43	32.61	22.45	23.46
Ah3R1	31.64	46.88	16.79	35.45	17.41	37.70	21.62	30.30

Tabla 29: EERs (%) de los sistemas *front-end* y *back-end* para experimentos inter-Base de Datos.

se tiene. Así, en nuestro caso de ahora, los vectores de parámetros del sistema *back-end* tienen 40 valores. O en otras palabras, el espacio de los *Anchor Models* es de de dimensión 40.

Cuando se realizaban experimentos *Intra*-Base de datos, el espacio de dimensión de los *Anchor Models* era de 22, 10 y 8 para cada base de datos *SUSAS Simulated*, *SUSAS Actual* y *Ah3R1* respectivamente.

Como ya se vio en el capítulo 3.7.4, AMF crea un nuevo vector de parámetros a partir de los resultados de los subsistemas *front-end* SVM con estadísticos y GMM-SVM. Cuanto mayor número de subsistemas se fusionen para crear este nuevo vector de parámetros mayor será la dimensión del mismo y por lo tanto según lo visto antes, mejores resultados obtendrá. Así, se va a realizar un nuevo AMF a partir de los resultados de SVM con estadísticos, GMM-SVM y además la fusión suma de ambos como se ve en la Figura 52. El nuevo vector de parámetros $\vec{S}_{x,m}$ está formado con las puntuaciones de los 3 subsistemas tendrá 60 coeficientes.

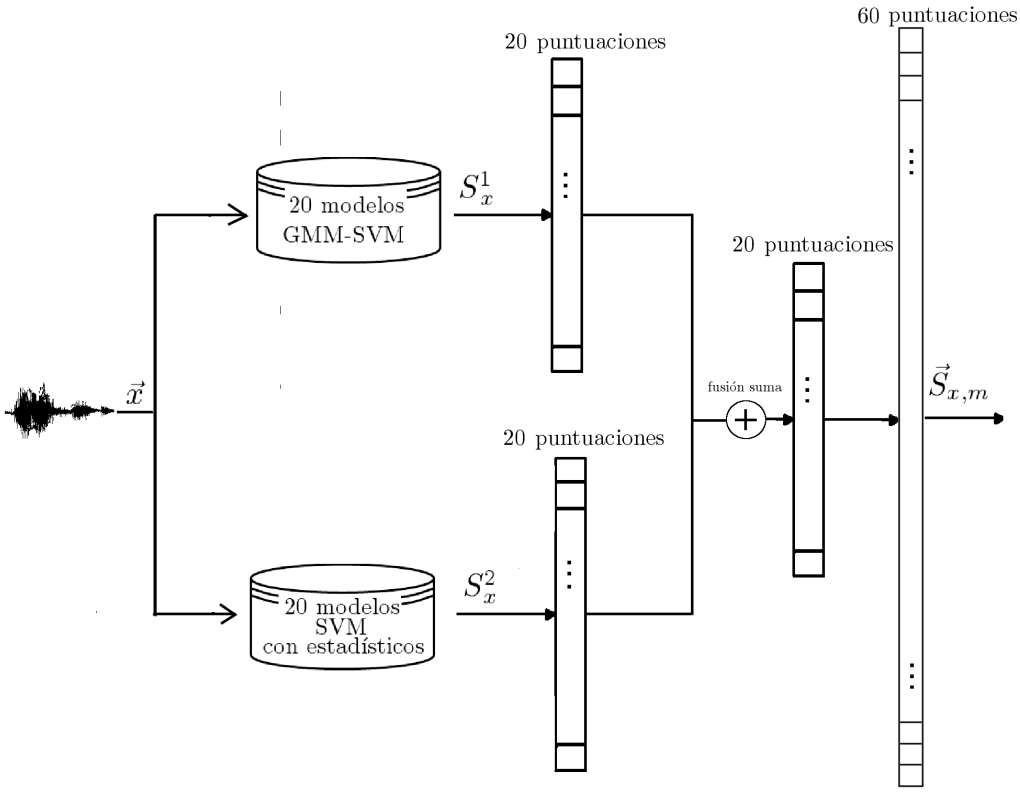

Figura 52: Uso de las puntuaciones de dos subsistemas *front-end* y de la fusión suma para conformar el nuevo sistema *back-end* de AMF.

Los resultados de este nuevo AMF formado por la fusión de 3 subsistemas aparecen en la Tabla 30 al igual que los del anterior AMF en lo que se fusionaban 2 subsistemas.

Como se ve en la Tabla 30, la nueva AMF consigue mejorar ligeramente los resultados para

	AMF (fusión 2 subsistemas)		**AMF (fusión 3 subsistemas)**	
	EER_{global}	EER_{medio}	EER_{global}	EER_{medio}
Simulated	27.74	28.80	25.96	26.29
Actual	22.45	23.46	21.89	23.01
Ah3R1	21.62	30.30	21.76	30.72

Tabla 30: EERs (%) para los dos tipos de sistemas AMF.

SUSAS Simulated y *SUSAS Actual* , pero no para *Ah3R1* en los que empeoran un poco.

Añadiendo la fusión suma a los AMF no se consiguen mejorar considerablemente los resultados. Esto es debido a que los resultados de la fusión suma son combinación de los otros dos subsistemas *front-end* de SVM con estadísticos y GMM-SVM, y por lo tanto no se añade mucha más información.

5.2. Pruebas y Resultados dependientes de locutor

Con experimentos dependientes de locutor eliminamos la variabilidad inter locutor pues los modelos serán entrenados con datos de un sólo locutor.

Se van a presentar y analizar los resultados para *SUSAS Simulated y Actual* obtenidos para los dos subsistemas *front-end* y su fusión suma y para el sistema *back-end* de AMF.

Parte de los resultados de estos experimentos han sido recogidos en [25] y aceptados para el congreso internacional *Interspeech 2009*.

SUSAS Simulated

Para la base de datos *SUSAS Simulated* la distribución de los locutores es la que aparece en la Tabla 31:

Etapa	**Locutores**
Development	*g1,b1,n1*
Entrenamiento y Test	*g2,b2,n2,g3,b3,n3*

Tabla 31: Distribución de locutores para experimentos dependientes de locutor en *SUSAS Simulated*.

No se separan unos locutores para entrenar los modelos y otros para evaluar, sino que datos de un mismo locutor los usamos tanto para entrenar modelos como para evaluarlos.

• *SUSAS Simulated* - Sistemas *front-end*: SVM con estadísticos, GMM-SVM y fusión suma

Se entrenan modelos por cada emoción y locutor. Además se hace *cross validation*, es decir, se entrenan modelos de la forma $\vec{w}_{loc\_emoc\_notWordX}$ donde *loc* es cada uno de los 6 locutores de entrenamiento, *emoc* es cada una de las 11 emociones de *SUSAS Simulated* y *notWordX* significa que dicho modelo es entrenado con locuciones de entre el conjunto de las 35 palabras menos la palabra *WordX*. Así, por ejemplo, el modelo $\vec{w}_{f2\_a\_notBreak}$ será entrenado

con locuciones del locutor *f2*, de la emoción *a (angry)* y con todas menos la palabra *break*.

Por lo tanto, el número de modelos para *SUSAS Simulated* para cada subsistema *front-end* (SVM con estadísticos y GMM-SVM) es de 11 emociones * 6 locutores * 35 palabras = 2310 modelos.

Una vez se han entrenado los 2310 modelos pasamos a la tarea de evaluación. Así, una locución de test como por ejemplo *break1.n2c5* se evalúa frente a los siguientes 11 modelos: $\vec{w}_{n2\_emoc\_notBreak}$ donde *emoc* es cada una de las 11 emociones. Con validación cruzada (*cross validation*) se consigue entonces que no se utilicen las mismas locuciones para entrenar y evaluar.

La Figura 53 muestra un esquema de la manera de entrenar y evaluar descrita anteriormente para el sistemas *front-end* de SVM con estadísticos. La mecánica para el sistema GMM-SVM es la misma pero por cada locución de entrenamiento y test se crea un modelo GMM mediante la adaptación del UBM generado con los datos de los locutores *g1,b1 y n1*.

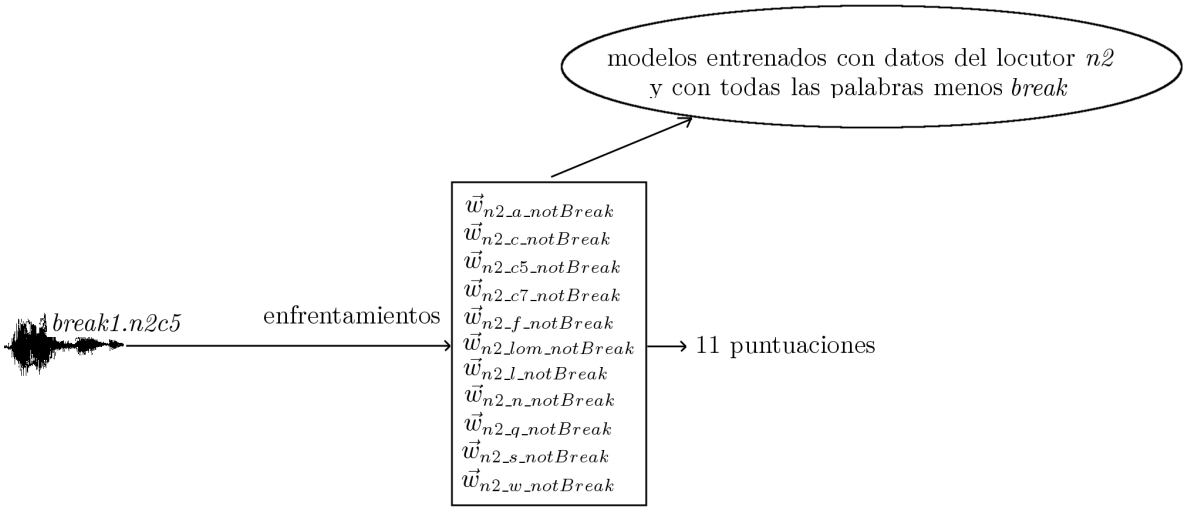

Figura 53: Esquema de la evaluación de las pruebas dependientes de locutor para '*SUSAS Simulated* - SVM con estadísticos'.

Para ambos subsistemas se van a ajustar los siguientes parámetros:

- Optimización variable M número de gaussianas (sólo subsistema GMM-SVM)
- Optimización variable *coste* de entrenamiento
- T-normalización de puntuaciones

No se ha incluido la normalización de los parámetros prosódicos como una tarea a ajustar ya que se va a usar la configuración que mejores resultados dio para los experimentos independientes de locutor. Así, para el subsistema de SVM con estadísticos se normalizará únicamente el vector de energía $\vec{e}$ mientras que para el de GMM-SVM se normalizará tanto el vector de energía $\vec{e}$ como el de su velocidad $\vec{\Delta}_e$.

En primer lugar se ajusta el valor del *coste* del clasificador SVM para ambos subsistemas. La Tabla 32 y la Figura 54 ofrecen los resultados en forma de tasas de error y DCF_{min} para varios valores de *coste* para el subsistema de SVM con estadísticos.

Según la Tabla 32, se aprecia que a medida que aumenta el *coste* se obtienen menores tasas de error, sin embargo no es apropiado establecer un valor de *coste* muy alto pues el tiempo de entrenamiento se dispara. Es por eso por lo que no hemos realizado la prueba para un valor

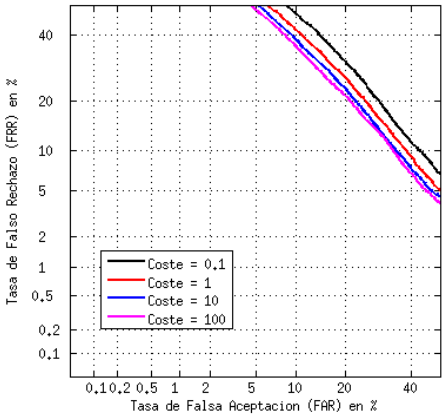

Figura 54: Curvas DET del sistema '*SUSAS Simulated* - SVM con estadísticos' para diferentes *costes* de entrenamiento.

Coste	**EER global(%)**	**DCF**$_{min}$	**EER medio (%)**
0.1	24.99	0.0887	21.22
1	22.80	0.0830	19.20
10	21.40	0.0805	17.99
100	20.69	0.0797	17.30

Tabla 32: Resultados dependiendo del valor del *coste* para '*SUSAS Simulated* - SVM con estadísticos'.

de 1000. Por lo tanto, nos quedamos con un *coste* de 100 como valor óptimo. La Figura 54 también deja claro que ésta es la mejor opción.

Para el subsistema de GMM-SVM se optimiza el número de gaussianas M. Tras varias pruebas probando con valores potencia de 2 se tomó el caso de 256 gaussianas pues es el que mejor resultados obtenía para un valor de *coste* fijo.

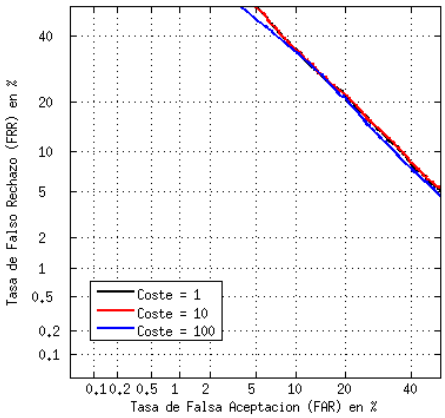

Figura 55: Curvas DET para '*SUSAS Simulated* - GMM-SVM' variando el *coste*.

Ahora se ajustan el *coste* para el subsistema GMM-SVM manteniendo fijo M a 256. Los resultados aparecen en forma gráfica en la Figura 55 y numéricamente en la Tabla 33.

Al igual que para el subsistema de SVM con estadísticos, a medida que el *coste* es mayor, se obtienen mejores resultados. Así, se elege el valor de *coste* de 100 pues es la opción que ofrece

Coste	EER global(%)	DCF_{min}	EER medio (%)
1	20.86	0.0857	17.32
10	20.94	0.0848	17.28
100	20.50	0.0755	17.17

Tabla 33: Resultados para '*SUSAS Simulated* - GMM-SVM' para varios *costes*.

mejores resultados sin que el tiempo de entrenamiento se dispare.

La Tabla 34 es la configuración final de tanto el subsistema SVM con estadísticos como del GMM-SVM una vez se ha hecho T-normalización de las puntuaciones finales.

	Norm $\vec{u}_p$	M	coste	*T-norm*	EER_{glob}	DCF_{min}	EER_{med}
SVM con estadísticos	$\vec{e}$	-	100	sí	20.84	0.0820	16.13
GMM-SVM	$\vec{e}$ y $\vec{\Delta}_e$	256	100	sí	18.24	0.0733	15.29
fusión suma					15.63	0.068	12.15

Tabla 34: Configuración y resultados optimizados para '*SUSAS Simulated* - SVM con estadísticos y GMM-SVM'.

Las curvas DET de tanto los 2 subsistemas por separado como la de la fusión suma aparecen en la Figura 56. Como dicha fusión suma se ha de realizar sobre resultados T-normalizados, tomaremos los datos de la Tabla 34 como configuración de nuestros sistemas.

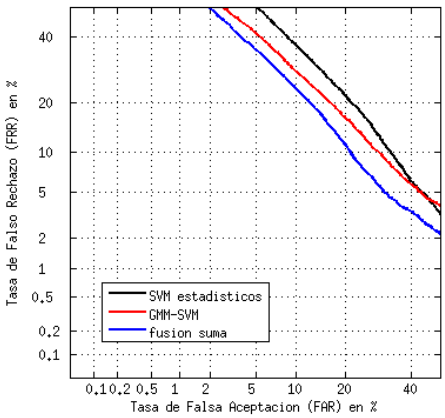

Figura 56: Curva DET de '*SUSAS Simulated* - SVM con estadísticos, GMM-SVM y fusión suma'.

- ***SUSAS Simulated* - Fusión de *Anchor Models* (AMF)**

Las puntuaciones de cada locución de test (ejemplo: *break1.n2c5*) obtenidas tras evaluarla frente a los 11 modelos de la forma $\vec{w}_{n2_emoc_notBreak}$ de cada uno de los subsistemas GMM-SVM y SVM con estadísticos se concatenan para conformar un nuevo vector de parámetros. Dichas puntuaciones serán las correspondientes a la configuración que en cada caso ha dado los mejores resultados [Tabla 34]. Dicho vector $\vec{S}_{break1.n2c5}$ tendrá 22 valores [Ver Figura 57].

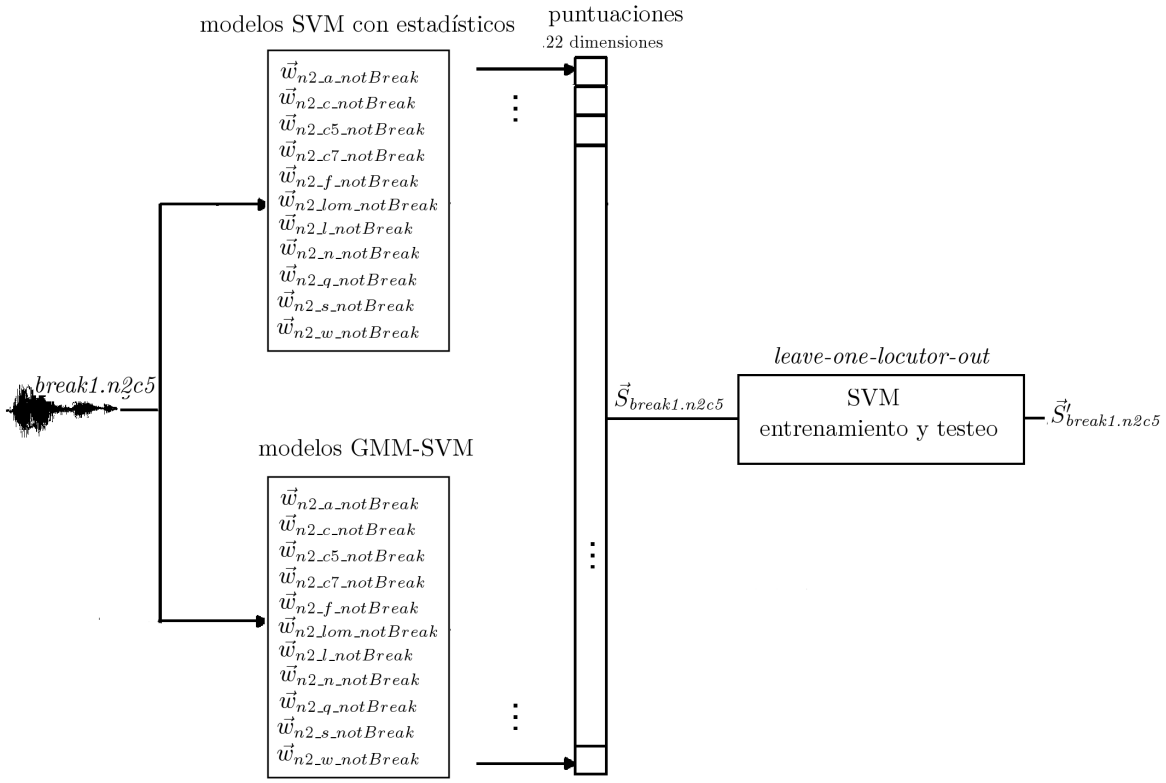

Figura 57: Esquema de las pruebas dependientes de locutor para '*SUSAS Simulated* - AMF'.

Una vez se tiene por cada locución de test $\vec{x}$ un nuevo vector de parámetros $\vec{S}_x$, éstos se utilizan como vectores de entrada a un clasificador SVM.

Estos nuevos modelos de la forma $\vec{w}'_{loc_emoc_notWordX}$ del sistema *back-end* se crean de igual manera que se crearon en los subsistemas *front-end*. Es decir, se crean con datos del locutor *loc* y emoción *emoc* y con todas la palabras menos la *WordX*. Y se evalúan con las locuciones de ese mismo locutor y que sean de la palabra *WordX*.

En la Figura 58 se representan un conjunto de curvas DET para varios valores de la variable coste del clasificador *back-end* SVM. Y en la Tabla 35 valores numéricos de tasas de error y DCF_{min}.

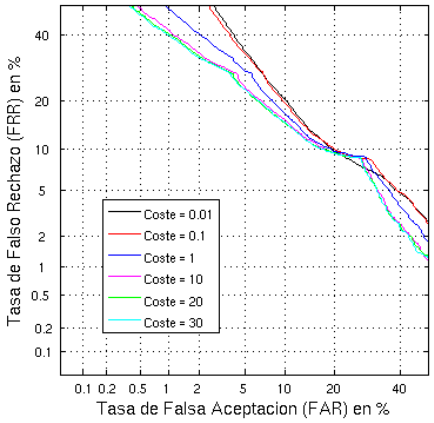

Figura 58: Curvas DET para '*SUSAS Simulated* - AMF' y varios *costes*.

Coste	EER global(%)	DCF_{min}	EER medio (%)
0.01	13.99	0.0751	9.30
0.1	13.82	0.0725	9.67
1	13.12	0.0583	9.02
10	12.72	0.0508	8.56
20	12.50	0.0499	8.26
30	12.50	0.0493	8.58

Tabla 35: Resultados dependiendo del *coste* para '*SUSAS Simulated* - AMF'.

Tanto para *coste* 20 como 30 se alcanza la más baja y mínima tasa de error global. Sin embargo la que optimiza los resultados es el *coste* 20 pues logra una tasa de error media 3 décimas inferior.

Una vez vistos los resultados para AMF, se compararán dichos resultados con los de la fusión suma de los subsistemas *front-end*.

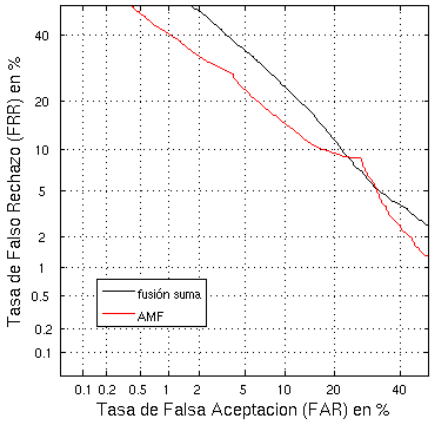

Figura 59: Curvas DET para '*SUSAS Simulated* - fusión suma y AMF'.

Con AMF se consigue una EER_{global} de 12.50 % mientras que la fusión de los sistemas *front-end* obtenía un 15.63 %. Es decir, se reduce en más de 3 puntos la tasa de error media. En la Figura 59 se representa la curva DET para la fusión suma del sistema SVM de estadísticos con el sistema GMM-SVM y la curva DET para el sistema de AMF.

Por último, la Tabla 36 analiza los EER_{medio} por emoción de tanto la fusión suma de los dos sistemas *front-end* como del sistema *back-end* de AMF. La última columna corresponde con la mejora relativa (M.R. en %) que ofrece éste último sistema con respecto al primero.

En la Figura 60 aparecen las curvas DET para cada una de las emociones de *SUSAS Simulated* para la fusión suma de los subsistemas *front-end*.

Mientras que en la Figura 61 representa las curvas DET para cada una de las emociones de *SUSAS Simulated* para AMF.

El rasgo más llamativo de la Tabla 36 es el estilo de habla *question* pues alcanza unas tasas de error muy bajas tanto para la fusión suma (2.2 %) como para AMF (1.08 %). Igualmente, los estilos *angry* y *loud* también obtienen porcentajes de error muy reducidos, por debajo del 10 %. Como viene siendo habitual, los estilos que ofrecen peores resultados son *cond50 y cond70*.

Emoción	EER (%) fusión suma	EER (%) AMF	M.R. (%)
angry	9.16	7.41	-19.1
clear	25.84	14.29	-44.7
cond50	23.40	20.90	-10.68
cond70	21.12	20.48	-3.03
fast	15.62	14.80	-5.25
lombard	14.13	9.92	-29.79
loud	10.09	4.88	-51.64
neutral	25.90	23.56	-9.03
question	2.22	1.08	-51.35
slow	13.53	12.96	-4.21
soft	10.39	7.73	-25.6

Tabla 36: EER (%) por emoción para '*SUSAS Simulated* - fusión suma y AMF'.

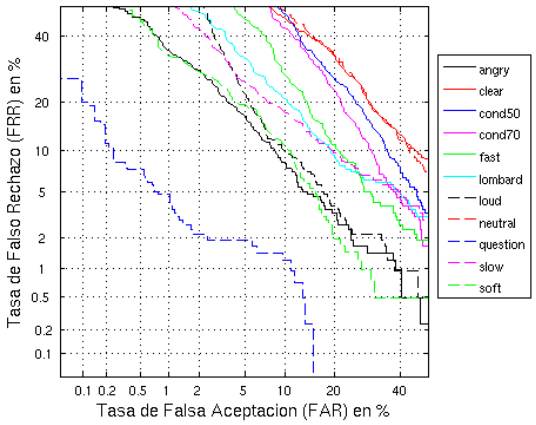

Figura 60: Curva DET para la fusión suma por emoción.

Como se vio en la Tabla 11, en los experimentos independientes de locutor el sistema de AMF parecía funcionar relativamente bien excepto para los estilos de habla *angry y loud*. Sin embargo estos son dos de los estilos de habla para los que AMF funciona mejor en experimentos dependientes de locutor [Ver Tabla 36].

En las Figuras 60 y 61 se representan gráficamente las curvas DET para la fusión suma y AMF por emoción respectivamente. Llama la atención la similitud entre ambas gráficas en cuanto al rendimiento relativo por emoción. Esto implica que ambas técnicas ofrecen resultados relativos y globales similares. Así, la curva de la emoción *question* (azul discontinua) es la que mejores tasas consigue y por ello más cercana al origen se encuentra. La siguen las curvas de *angry* (negra), *loud* (negra discontinua) y *soft* (verde discontinúa). También para ambas gráficas las curvas más alejadas del origen y por lo tanto las de peores resultados son las de los estilos *clear* (roja), *cond50* (azul) y *cond70* (rosa).

SUSAS Actual

La distribución de los locutores aparece en la Tabla 37.

• ***SUSAS Actual* - Sistemas *front-end*: SVM con estadísticos, GMM-SVM y fusión suma**

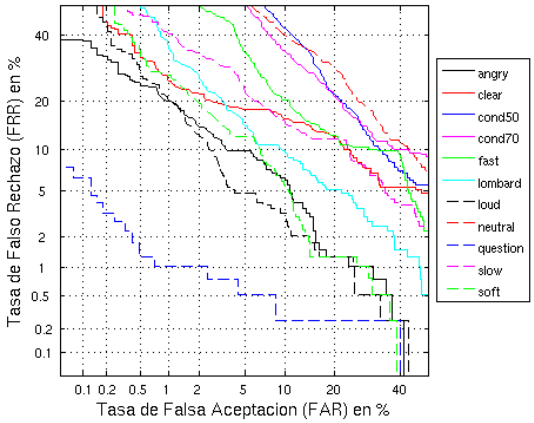

Figura 61: Curvas DET por emoción para '*SUSAS Simulated* - AMF'.

Etapa	Locutores
Development	*f1,m1*
Entrenamiento y Test	*f2,m2,f3,m3,m4*

Tabla 37: Distribución de locutores para experimentos dependientes de locutor en *SUSAS Actual* .

Al igual que en *SUSAS Simulated*, se entrenan modelos por locutor y emoción y además se implementa *Leave-One-Word-Out*. Los modelos son de la forma $\vec{w}_{loc\_emoc\_notWordX}$ donde *loc* es cada uno de los 5 locutores de entrenamiento, *emoc* es cada una de las 5 emociones de *SUSAS Actual* y *WordX* es la palabra que no se usará para el entrenamiento de ese modelo. Por ejemplo, el modelo $\vec{w}_{m4\_f\_notHello}$ será entrenado con locuciones del locutor *m4*, de la emoción *f (freefall)* y con todas menos la palabra *hello*. El número de modelos para *SUSAS Actual* por cada subsistema *front-end* (SVM con estadísticos y GMM-SVM) es 5 emociones * 5 locutores * 35 palabras = 875modelos.

Una locución de test como por ejemplo *hello2.m4f* se evalúa frente a los siguientes 5 modelos: $\vec{w}_{m4_emoc_notHello}$ donde *emoc* es cada una de las 5 emociones.

Los datos de los locutores *f1 y m1* se usan para generar el modelo UBM para el subsistema GMM-SVM.

Para ambos subsistemas *front-end* (SVM con estadísticos y GMM-SVM) se van a optimizar los siguientes parámetros:

- Optimización variable M número de gaussianas (sólo subsistema GMM-SVM)
- Optimización variable coste de entrenamiento
- Normalización de los vectores de parámetros prosódicos

Todos los resultados que aparecen para estos tipos de experimentos serán tras haber hecho T-normalización de puntuaciones pues se ha visto que siempre supone una mejora sobre el sistema.

Los parámetros del subsistema GMM-SVM serán los primeros en ser ajustados. Se empieza con la normalización de los vectores prosódicos manteniendo fijo el número de gaussianas M a 32 y valor de *coste* de 1. Con ello, se obtienen los valores de EER global de la Tabla 38.

Normalización $\vec{u}_p$	EER global(%)
no	19.20
$\vec{e}$	20.30
$\vec{e}$ y $\vec{\Delta}_e$	20.20
$\vec{p}$	21.55
$\vec{p}$ y $\vec{\Delta}_p$	23.68
$\vec{e}$, $\vec{\Delta}_e$, $\vec{p}$ y $\vec{\Delta}_p$	23.78

Tabla 38: EER global dependiendo de los vectores de parámetros prosódicos normalizados para '*SUSAS Actual* - GMM-SVM'.

El valor de M es el siguiente en ser ajustado. Para esta tarea se mantienen los vectores de parámetros prosódicos originales, es decir, sin normalizar pues según la Tabla 38 es la que mejores resultados ofrece. La Tabla 39 tiene los EER globales para distintos valores de M.

M	EER global(%)
8	52.50
16	17.27
32	19.20
64	22.70

Tabla 39: EER global para '*SUSAS Actual* - GMM-SVM' dependiendo del número de gaussinas.

Por último, se varía la variable *coste* del clasificador SVM habiendo usado 16 gaussianas y no normalización de los vectores de parámetros. Es decir la configuración que ofrece mejores resultados. Estos resultados los podemos ver en la Tabla 40. Los resultados se optimizan con un valor de *coste* de 10.

Coste	EER global(%)
1	17.27
10	15.9
100	15.96

Tabla 40: EER global para '*SUSAS Actual* - GMM-SVM' dependiendo del *coste*.

Una vez ajustado el subsistema de GMM-SVM es el turno de optimizar el de SVM con estadísticos. Para este subsistema únicamente ajustaremos las variables de *coste* y normalización de los vectores prosódicos. Así, en la Tabla 41 aparecen los EER globales para varias normalizaciones.

Normalizando únicamente el vector $\vec{e}$ de energías se consiguen los mejores tasas de error. [Ver Tabla 41].

La Tabla 42 ofrece los resultados para varios valores de *coste* habiéndonos normalizado anteriormente el vector prosódico de energías. El valor de 100 de *coste* es el que ofrece mejores resultados de entre los tres que hemos probado. No se han probado valores más altos pues retardaban en exceso los tiempos de entrenamiento de los modelos.

Con todo esto, la Tabla 43 es la configuración final de tanto el subsistema SVM con estadísticos como del GMM-SVM una vez se ha hecho T-normalización de las puntuaciones finales:

Normalización $\vec{u}_p$	EER global(%)
no	21.86
$\vec{e}$	19.86
$\vec{e}$ y $\vec{\Delta}_e$	19.97
$\vec{p}$	21.04
$\vec{p}$ y $\vec{\Delta}_p$	21.80
$\vec{e}$, $\vec{\Delta}_e$, $\vec{p}$ y $\vec{\Delta}_p$	26.45

Tabla 41: EER global para '*SUSAS Actual* - SVM con estadísticos' según los vectores prosódicos normalizados.

Coste	EER global(%)
1	20.9
10	19.86
100	18.64

Tabla 42: EER global dependiendo del *coste* para '*SUSAS Actual* - SVM con estadísticos'.

Las curvas DET de tanto los 2 subsistemas por separado como la de la fusión suma aparecen en la Figura 62.

Como suele suceder, la fusión suma consigue mejorar los resultados de los subsistemas individuales. Este caso en concreto, consigue bajar en torno a 1 punto las tasas de error del subsistema GMM-SVM que es el mejor de los 2.

- ***SUSAS Actual* - Fusión de *Anchor Models* (AMF)**

Las puntuaciones de cada locución de test (ejemplo: *hello2.m4f*) obtenidas tras evaluarla frente a los 5 modelos de la forma $\vec{w}_{m4_e moc_n otHello}$ de cada uno de los subsistemas GMM-SVM y SVM con estadísticos para conformar un nuevo vector de parámetros. Este vector $\vec{S}_{hello2.m4f}$ tendrá 10 valores. Como siempre, estos nuevos vectores de parámetros $\vec{S}_x$ se usan como entrada a un clasificador SVM.

Estos nuevos modelos de la forma $\vec{w}'_{loc_e moc_n otWordX}$ del sistema *back-end* se crean con datos del locutor *loc* y emoción *emoc* y con todas la palabras menos la *WordX*. Y se evalúan con las locuciones de ese mismo locutor y que sean de la palabra *WordX*.

La Figura 63 contiene una serie de curvas DET para varios valores de la variable *coste* del clasificador *back-end* SVM. Y en la Tabla 44 valores numéricos de tasas de error y DCF_{min}.

	Norm $\vec{u}_p$	**M**	**coste**	***T-norm***	$\textbf{EER}_{global}$	$\textbf{DCF}_{min}$	$\textbf{EER}_{medio}$
SVM estadísticos	$\vec{e}$	-	100	sí	18.64	0.0902	17.05
GMM-SVM	no	16	10	sí	15.90	0.0816	11.64
fusión suma					15.02	0.0746	10.29

Tabla 43: Configuración y resultados optimizados para '*SUSAS Actual* - SVM con estadísticos y GMM-SVM'.

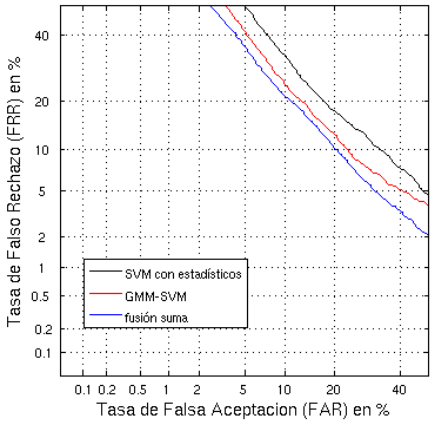

Figura 62: Curvas DET para '*SUSAS Actual* - SVM con estadísticos, GMM-SVM y fusión suma'.

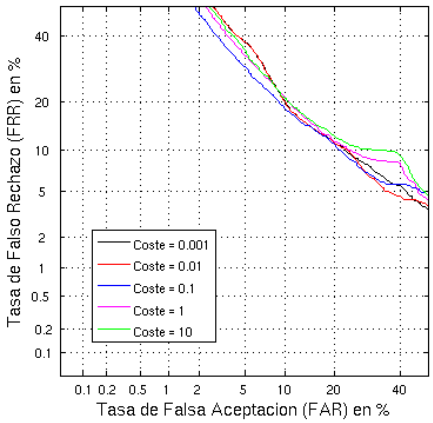

Figura 63: Curvas DET para '*SUSAS Actual* - AMF' y varios *costes*.

Coste	**EER global(%)**	**DCF$_{min}$**	**EER medio (%)**
0.001	14.44	0.0730	13.18
0.01	14.47	0.0729	12.17
0.1	14.39	0.0669	11.99
1	14.89	0.0725	12.38
10	15.17	0.0739	12.38

Tabla 44: Resultados para varios *costes* para '*SUSAS Actual* - AMF'.

Se elige el valor de *coste* que mejores resultados ofrece, es decir, el de 0.1.

Una vez vistos los resultados para AMF, se comparan dichos resultados con los de la fusión suma de los subsistemas *front-end* [Ver Figura 64].

Con AMF se consigue una EER$_{global}$ de 14.39 % mientras que la fusión de los sistemas *front-end* obtenía un 15.02 %. Entonces, AMF reduce en unas décimas la tasa de error global. Sin embargo si ahora se analizan los valores de EER$_{medio}$ se aprecia que AMF alcanza un 11.99 % por un 10.29 % de la fusión. Es decir, ahora AMF no mejora los resultados que ofrece la fusión suma. Se concluye que para *SUSAS Actual* el comportamiento de los sistemas *front-end* es muy similar al sistema *back-end* de AMF.

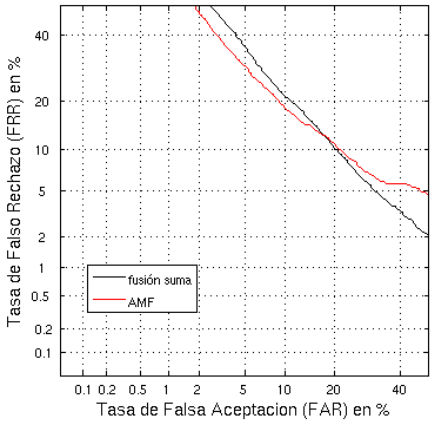

Figura 64: Curvas DET para '*SUSAS Actual* - fusión suma y AMF'.

Por último, la Tabla 45 analiza los EER_{medio} por emoción de tanto la fusión suma de los dos sistemas *front-end*, como del sistema *back-end* de AMF, como de la mejora relativa de este último sobre el primero.

Emoción	**EER (%) fusión suma**	**EER (%) AMF**	**M.R. (%)**
neutral	15.23	19.74	+29.61
medst	22.79	17.42	-23.56
hist	19.85	19.34	-2.08
freefall	20.97	18.43	-12.11
scream	5.72	6.37	+11.36

Tabla 45: EER (%) por emoción para '*SUSAS Actual* - fusión suma y AMF'.

En la Figura 65 aparecen las curvas DET para cada una de las emociones de *SUSAS Actual* para la fusión suma de los subsistemas *front-end*.

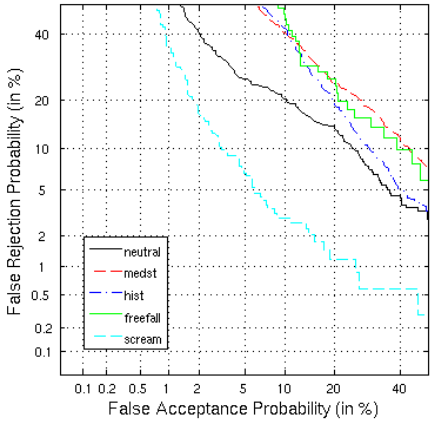

Figura 65: Curvas DET por emoción para '*SUSAS Actual* - fusión suma'.

Mientras que en la Figura 66 representa las curvas DET para cada una de las emociones de *SUSAS Actual* para AMF.

El estilo de habla que con diferencia ofrece mejores resultados para *SUSAS Actual* es según

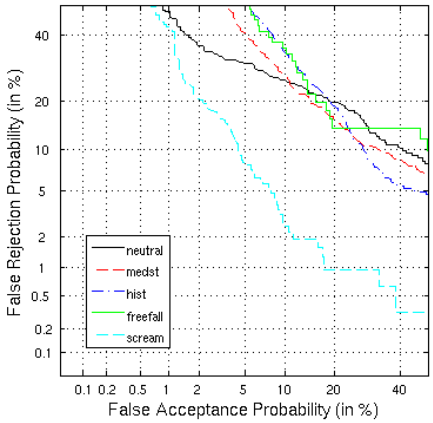

Figura 66: Curvas DET por emoción para '*SUSAS Actual* - AMF'.

la Tabla 45, *scream*. Alcanza unas tasas de error muy bajas tanto para la fusión suma (5.72 %) como para AMF (6.37 %). Los demás (*neutra, medst, hist y freefall*) obtienen resultados similares en torno al 15 %.

El sistema de AMF parece funcionar mejor para aquellas emociones que peor resultados obtienen. Así, para *medst, hist y freefall* la técnica de AMF mejora con respecto a la fusión suma, sobre todo para el estilo *medst* [Ver Tabla 45].

En las Figuras 65 y 66 se representan gráficamente las curvas DET para la fusión suma y AMF por emoción respectivamente. Lo que más llama la atenció a primera vista para ambas gráficas es que la emoción *scream* es con diferencia la que menores tasas de error produce.

Tras analizar los resultados dependientes de locutor, se presenta en la Tabla 46 una recopilación de los mejores EER medio para *SUSAS Simulated y Actual*.

base de datos	***front-end/back-end***	**EER medio (%)**
SUSAS Simulated	front-end(fusión suma)	12.15
	back-end (AMF)	8.26
SUSAS Actual	front-end(fusión suma)	10.29
	back-end (AMF)	11.99

Tabla 46: EER_{medio} (%) para las 3 bases de datos para experimentos dependientes de locutor.

Viendo la Tabla 28 de resumen, el sistema de fusión suma es más robusto para la base de datos *SUSAS Actual* mientras que AMF lo es para *SUSAS Simulated*.

Ah3R1

En un principio se intentaron realizar experimentos dependientes de locutor para la base de datos *Ah3R1*. Sin embargo los resultados obtenidos fueron mucho peores al del resto de bases de datos para este mismo tipo de experimentos. La razón de dichos malos resultados posiblemente sea que dada la escasez de datos de entrenamiento por cada locutor y emoción que ofrece *Ah3R1*, no se consigue entrenar modelos correctamente adaptados a cada locutor.

Se espera que progresivamente vayan estando disponibles diferentes versiones de esta base de

datos ofreciendo así más volumen de información por locutor. Si así sucede, se conseguirá más robustez en experimentos dependientes de locutor para esta base de datos.

6

Conclusiones y Trabajo futuro

6.1. Conclusiones

Este trabajo se ha focalizado en la evaluación y desarrollo de sistemas para el reconocimiento automático de emociones en el habla. Nuestros resultados son similares a los obtenidos en el estado del arte, incluso en algunos casos son considerablemente mejores.

Parte de esta evaluación consistió en la realización de experimentos en los que se examinó desde la influencia de ciertas variables en el comportamiento del sistema, hasta su rendimiento tras la implementación de distintos tipos de normalizaciones de tanto los vectores paramétricos como de las puntuaciones.

Las variables ajustadas fueron por un lado el coste del entrenamiento y el número de mezclas gaussianas M. El coste no ha seguido un comportamiento regular en cuanto a los resultados, la única influencia ha sido sobre el tiempo empleado en el entrenamiento de los modelos. Aunque la elección de un valor alto de M implica una mejor adaptación a los resultados, no siempre ha sido posible por la escasez de datos disponibles. Por otro lado, se realizaron dos tipos de normalizaciones. Una, la *T-norm*, sobre las puntuaciones, la cual mostró siempre una leve mejora en el comportamiento del sistema. La otra, sobre los vectores de parámetros prosódicos, comprueba que los únicos vectores sobre los que su normalización logra mejorar los resultados son el vector de energías $\vec{e}$ y su velocidad $\vec{\Delta}_e$.

Para experimentos Intra-Base de datos, la nueva técnica de AMF logra mejorar los resultados de la fusión suma en todos los casos menos para experimentos independientes de locutor sobre la base de datos *SUSAS Actual*. Mientras que en experimentos Inter-Base de datos AMF siempre supera a la fusión suma. Para estos últimos experimentos existe un mayor número de modelos *Cohorte* y por lo tanto la dimensión de los *Anchor Models* es mayor. Una mayor dimensión implica que sea más probable la discriminación entre emociones que para un número pequeño de emociones como ocurre en los experimentos Intra-Base de datos.

La teoría anterior se vuelve a comprobar si comparamos los resultados para el sistema AMF formado a partir de los resultados de los dos subsistemas *front-end* y el otro, formado a partir de los dos subsistemas *front-end* además de su fusión suma. Este último AMF, al tener vectores de mayor dimensión obtiene inferiores tasas de error que el primero. Por ejemplo para la base de datos *SUSAS Simulated* se pasa de un 28.8 % de EER medio a un 26.29 %.

Las mejoras relativas más importantes de AMF sobre la fusión suma se logran en la base de datos *SUSAS Simulated* pues al tener más emociones que el resto de bases de datos, es donde más dimensiones de los *Anchor Models* se tiene.

No todas las emociones se comportan de igual manera. Así, en pruebas independientes de locutor, AMF mejora menos o empeora más con respecto a la fusión suma en aquellas emociones que se caracterizan por tener valores altos de energía y de su variación como son *angry*, *loud* para *SUSAS Simulated*, *scream* para *SUSAS Actual* y *exaltado* para *Ah3R1*.

También, las emociones que menores tasas de error ofrecen son aquellas que se caracterizan por grandes variaciones de energía y pitch ya que son justamente estos dos parámetros con los que hemos caracterizado la señal de voz. Éstas son: *angry* y *question* en *SUSAS Simulated*, *scream* en *SUSAS Actual* y *exaltado* en *Ah3R1*. Así, por ejemplo, la emoción *scream* presenta un EER cercano al 5 % o mejor aún, la emoción *question* llega a alcanzar el 1.08 % en experimentos dependientes de locutor.

Cabe destacar los resultados dependientes de locutor alcanzados mediante la fusión del sistema SVM con estadísticos y el sistema híbrido GMM-SVM y mediante la fusión de *Anchor Models* para la base de datos *SUSAS Simulated*. Los resultados presentan un EER medio del

12.15 % y 8.26 % respectivamente. Estos resultados sitúan a nuestros sistemas en una muy buena posición en el estado del arte actual.

Los resultados de los experimentos dependientes de locutor son considerablemente mejores que los independientes de locutor pues eliminan la variabilidad de locutor. De esta manera, al tratar con datos de un solo locutor, éstos abarcan mucha menos diversidad de habla que si manéjasemos datos de todos los locutores.

La base de datos que ofrece mejores resultados es *SUSAS Simulated* pues, aunque éstos son similares a los obtenidos sobre *SUSAS Actual*, se comprende de 11 emociones por solo 5 de *SUSAS Actual* y por lo tanto la tarea de reconocimiento de emociones se hace más difícil. Sin embargo, los resultados más realísticos serían los obtenidos sobre *SUSAS Actual* y *Ah3R1* pues contienen datos reales y espontáneas mientras que los *SUSAS Simulated* están posiblemente exagerados.

Por lo general los resultados son muy satisfactorios si los comparamos con el estado del arte actual, más aún si tenemos en cuenta que el reconocimiento de emociones es un campo nuevo en el grupo ATVS.

Los resultados obtenidos en este proyecto han dado lugar a dos publicaciones aceptadas y a la espera de ser publicadas en congresos internacionales:

• Lopez-Moreno, I., Ortego-Resa C., Gonzalez-Rodriguez J., Ramos D. , "Speaker dependen emotion recognition using prosodic supervectors", 2009.

• Ortego-Resa C., Lopez-Moreno, I., Gonzalez-Rodriguez J., Ramos D. , "Anchor model fusion for emotion recognition in speech", 2009.

6.2. Trabajo futuro

A partir del presente trabajo, existen varias líneas de investigación en el campo de las emociones en el habla. Una de ellas sería buscar el tipo de parametrización óptima para la discriminación entre emociones. Un primer paso sería la combinación de parámetros prosódicos y acústicos así como añadir rasgos de acceleración a los vectores prosódicos de energía y pitch.

También resulta interesante aplicar las técnicas de reconocimiento de emociones para tareas de reconocimiento de locutor a través de voz emocional o reconocimiento de habla emocional. Estas tareas aunque no consisten explícitamente en clasificar emociones, sí requieren el uso de sus técnicas.

Por último, otra línea futura de trabajo sería añadir el entrenamiento de los modelos SVM basados en regresión, *épsilon-SVR*. Esta es una de las investigaciones más importantes llevadas a cabo en el campo de reconocimiento de locutor y que podría migrar a reconocimiento de emociones logrando buenos resultados.

Bibliografía

[1] Anil K. Jain and David Maltoni, *Handbook of Fingerprint Recognition*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[2] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods", *Speech Communication*, vol. 48, no. 9, pp. 1162 – 1181, 2006.

[3] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, "The det curve in assessment of detection task performance", in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 1895–1898.

[4] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42 – 54, 2000.

[5] Zhihong Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[6] Bjöern Schuller, Ronald Müeller, Benedikt Höernler, Anja Höethker, Hitoshi Konosu, and Gerhard Rigoll, "Audiovisual recognition of spontaneous interest within conversations", in *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, New York, NY, USA, 2007, pp. 30–37, ACM.

[7] D. G. Childers and Ke Wu, "Gender recognition from speech. part ii: Fine analysis", *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1841–1856, 1991.

[8] T. Bocklet, A. Maier, J.G. Bauer, F. Burkhardt, and E.Ñoth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines", in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 1605–1608.

[9] Carl E. Williams and KennethÑ. Stevens, "Emotions and speech: Some acoustical correlates", *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[10] J.H.L. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition", in *Speaker Classification (1)*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 108–137, Springer.

[11] J.H.L. Hansen, "Evaluation of acoustic correlates of speech under stress for robust speech recognition", Mar 1989, pp. 31–32.

[12] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, May 2004, vol. 1, pp. I–577–80 vol.1.

[13] C. Pereira, "Dimensions of emotional meaning in speech", 2000.

[14] Rodman R.D. Eriksson, E.J. and R.C. Hubal, "Emotions in speech: Juristic implications", in *Speaker Classification (1)*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 152–173, Springer.

[15] Daniel Neiberg1, Kjell Elenius, and Kornel Laskowski, "Emotion recognition in spontaneous speech using gmms", in *Interspeech 2006*, 2006.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[17] Bogdan Vlasenko, "Combining frame and turn-level information for robust recognition of emotions within speech", in *Interspeech 2007*.

[18] Navas E. Hernáez I. Luengo, I. and J. Sánchez, "Automatic emotion recognition using prosodic parameters", in *Interspeech 2005*.

[19] Vladimir Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[20] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee, "Emotion recognition by speech signals", in *EUROSPEECH-2003*, 2003, pp. 125–128.

[21] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez, "Automatic emotion recognition using prosodic parameters", in *EUROSPEECH 2005*.

[22] Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals", in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[23] Hao Hu, Ming-Xing Xu, and Wei Wu, "Gmm supervector based svm with spectral features for speech emotion recognition", in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–413–IV–416.

[24] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation", May 2006, vol. 1, pp. I–I.

[25] Ortego-Resa C. Gonzalez-Rodriguez J. Ramos D. Lopez-Moreno, I., "Speaker dependent emotion recognition using prosodic supervectors", 2009.

[26] V.Ñ. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[27] I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez, and D. T. Toledano, "Anchor-model fusion for language recognition", in *Proceedings of Interspeech 2008*, September 2008.

[28] Lopez-Moreno I. Gonzalez-Rodriguez J. Ramos D. Ortego-Resa, C., "Anchor model fusion for emotion recognition in speech", 2009.

[29] Hua Yu and Jie Yang, "A direct lda algorithm for high-dimensional data – with application to face recognition", *Pattern Recognition*, vol. 34, no. 10, pp. 2067 – 2070, 2001.

[30] Tin Lay Nwe, Say Wei Foo, and L.C. De Silva, "Classification of stress in speech using linear and nonlinear features", April 2003, vol. 2, pp. II–9–12 vol.2.

[31] J.H.L. Hansen and S.E. Bou-Ghazale, "Getting started with susas: a speech under simulated and actual stress database", in *EUROSPEECH-1997*, 1997, pp. 1743–1746.

[32] John H. L. Hansen, *SUSAS*, Linguistic Data Consortium, 1999.

[33] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-case database in spanish", in *Proceedings of Interspeech 2008*, September 2008, pp. 1493–1496.

[34] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "Ahumada: a large speech corpus in spanish for speaker characterization and identification", *Speech Communication*, vol. 31, pp. 255–264, June 2000.

[35] M. Grimm, K. Kroschel, and S.Ñarayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech", in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–1085–IV–1088.

[36] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.04) [computer program]", Ap 2009, http://www.praat.org/.

[37] Nello Cristianini, "Kernel methods for pattern analysis", in *ICTAI '03: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, Washington, DC, USA, 2003, p. .21, IEEE Computer Society.

[38] Ron Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", 1995, pp. 1137–1143, Morgan Kaufmann.

Glosario de acrónimos

- **AMF**: Anchor Model Fusion
- **ANN**: Artificial Neural Networks
- **DCF**: Detection Cost Function
- **EER**: Equal Error Rate
- **EM**: Expectation-Maximization
- **FA**: Falsa Aceptación
- **FR**: Falso Rechazo
- **GMM**: Gaussian Mixture Model
- **HMM**: Hidden Markov Model
- **LDA**: Linear discriminant analysis
- **LDC**: Linguistic Data Consortium
- **LFPC**: Low Frecuency Power Coefficients
- **MAP**: Maximum A Posteriori
- **MFCC**: Mel-Frequency Cepstral Coefficients
- **NIST**: National Institute of Standards and Technology
- **ROC**: Receiver Operating Curve
- **SDC**: Shifted Delta Cepstral
- **SRE**: Speaker Recognition Evaluation
- **SUSAS**: Speech Under Simulated and Actual Stress
- **SVM**: Support Vector Machine
- **T-norm**: Test Normalization
- **UBM**: Universal Background Model
- **VAD**: Voice Activity Detector
- **Z-norm**: Zero Normalization

A

Anexo: publicaciones

Publicaciones en congresos internacionales (aceptadas y a la espera de ser publicadas)

• Lopez-Moreno, I., Ortego-Resa C., Gonzalez-Rodriguez J., Ramos D. , "Speaker dependent emotion recognition using prosodic supervectors", 2009.

• Ortego-Resa C., Lopez-Moreno, I., Gonzalez-Rodriguez J., Ramos D. , "Anchor model fusion for emotion recognition in speech", 2009.

Speaker Dependent Emotion Recognition Using Prosodic Supervectors

Ignacio Lopez-Moreno, Carlos Ortego-Resa, Joaquin Gonzalez-Rodriguez and Daniel Ramos

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

ignacio.lopez@uam.es

Abstract

This work presents a novel approach for detection of emotions embedded in the speech signal. The proposed approach works at the prosodic level, and models the statistical distribution of the prosodic features with Gaussian Mixture Models (GMM) mean-adapted from a Universal Background Model (UBM). This allows the use of GMM-mean supervectors, which are classified by a Support Vector Machine (SVM). Our proposal is compared to a popular baseline, which classifies with an SVM a set of selected prosodic features from the whole speech signal. In order to measure the speaker inter-variability, which is a factor of degradation in this task, speaker dependent and speaker independent frameworks have been considered. Experiments have been carried out under the SUSAS subcorpus, including real and simulated emotions. Results shows that in a speaker dependent framework our proposed approach achieves a relative improvement greater than 14% in Equal Error Rate (EER) with respect to the baseline approach. The relative improvement is greater than 17% when both approaches are combined together by fusion with respect to the baseline.

Index Terms: emotion recognition, speaker inter-variability, supervectors, SVMs

1. Introduction

Emotion recognition from the speech signal is an increasingly interesting task in human-machine interaction, with diverse applications in the speech technologies field such as call centres, intelligent auto-mobile systems, speaker intra-variability compensation or entertainment industry [1]. Emotion recognition is generally stated as a multiclass classification problem, where a given speech utterance is classified among n emotions (classes). However, it is of interest to detect a given emotion in a speech segment, which justifies the use of a verification or detection approach described as follows: given a speech utterance and a target emotional state e from the whole n emotions set, the objective is to determine whether the dominant emotion that affect the speaker in the utterance is e or not. Thus, emotion detection is essentially a two-class problem, where the *target* class is true when e is the dominant emotion in the test utterance and the *non-target* class is true when it is not. The standard architecture in such scheme is to compute a similarity measure (a *score*) among an emotion model of e and the emotion in the test utteranace, which will be further compared to a threshold for detection.

Recognizing emotions from speech is essentially motivated from their nature: affective states caused by subjective judgements, memories and sensations frequently accompanied of physical and psychological changes of the well-being sensation. Thus humans can recognize emotions by the study of those changes of the neutral states, including the semantic level of the speech, non usual behaviours and decisions, as well as other not so high cognitive levels, commonly more capable to be learned by machines [2].

Unluckily, emotion recognition from speech is a difficult task, mainly because of two reasons. First, emotions does not manifest in the same way in different speakers, and therefore, inter-variability of speakers seriously affects the recognition process. Second, it is difficult to define the target emotions set because the limits among different emotions may not be clear for listeners in general, and several emotions from the considered set can be simultaneously in the same utterance, or even at the same moment in time. Despite the difficulty of the challenge, the research in the area has experimented an increase in the last years, which has motivated the availability of emotional labeled speech corpora. Most popular ones are FAU AIBO Emotion Corpus [3], SUSAS, EMO-DB, ISL meeting corpus, Danish Emotional Speech Database [4] and recently Ahumada III [5].

In this work, we present a novel method for emotions detection based on Gaussian Mixture Models (GMM) of short-term prosodic features, whose supervectors are further classified with Support Vector Machines (SVM). Moreover, we present results of the fusion of the proposed system with a baseline, based on a popular approach of modelling utterance-level prosodic features with SVM. We show that the proposed approach, namely prosodic SVM-GMM, models distances among complete join probability distributions of the prosodic features, and not only with some significant values, as happen with the baseline system. Moreover, the fusion of both systems significantly improves the performance of proposed approach, which indicates uncorrelated information among both methods. We evaluate the proposed system in a speaker-dependent and a speaker-independent scenario. Experiments are presented using the SUSAS database [6].

This work is organised as follows. The role of prosody and the proposed prosodic parametrization is described in Section 2. In Section 3, the proposed system is described in detail, as well as the baseline and the approach for fusion of both systems. Section 4 describes the experimental work which shows the adequacy of the approach. Finally, conclusions are drawn in Section 5.

2. Prosodic features for emotion recogniton

Many works had shown the relation between the variation of speaker prosody and the information of their emotional states [7]. Therefore prosodic features are often considered as input signals in many emotion recognition systems. Frequent prosodic features are the fundamental frequency (*pitch*), the energy and their velocity, also known as Δ features [8].

The proposed GMM-SVM approach in this work uses a prosodic feature extraction scheme in the following way: the audio signal is windowed every 10ms using a 40ms Hamming

window. For every window, energy and log pitch values are extracted (Fig.1) using Praat [9] toolbox. In vocal segments, velocity information is obtained as a difference between two consecutive windows. Using a voice activity detector (VAD), non-voiced segments are erased by accepting only those windows with pitch and energy values higher than a threshold. As a consequence, for every utterance u, the feature vector set consist of a set of $d = 4$ dimensional feature vectors, or streams (energy, pitch and their Δ features). It is possible to normalize each stream by subtracting its mean value. Energy and delta-energy normalization have been applied to the proposed GMM-SVM approach while only energy normalization for the baseline.

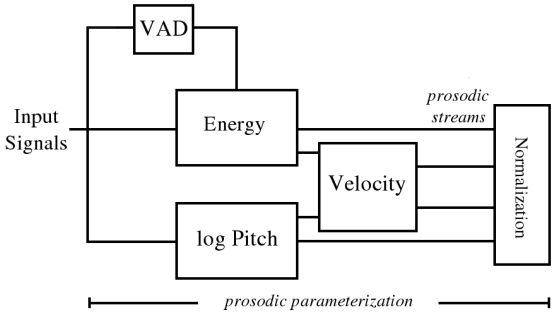

Figure 1: *Block diagram of the prosodic feature extraction module.*

3. A prosodic GMM-SVM approach for emotion detection

This section details the novel prosodic GMM-SVM system proposed in this paper, the baseline modelling scheme and the fusion approach for combining information from both systems.

3.1. Proposed approach

SVM-GMM supervectors have been previously used for emotion recognition at the spectral level of the speech in [10]. This technique also shows an excellent performance in speaker and language recognition. The main advantage of this proposed technique is that it is capable to summarize the whole probability density function (pdf) of the feature vectors in utterance u, into a single high-dimensionality vector known as a GMM supervector. This supervector is obtained by the concatenation of the vectors of means of a d-dimensional GMM model obtained from all the d-dimensional prosodic vectors in the utterance (Figure 2). The M-mixture GMM, is calculated as a Maximum a Posteriori Adaptation (MAP) from a Universal background Model (UBM), which is an standard M-mixtures GMM model, trained with a large amount of development data from all the emotional states available. Thus, the UBM aims at representing the emotion-independent statistical distribution of the features.

The GMM supervector can be considered as a kernel function $sv(u)$ that maps the prosodic features of u in a high-dimensional vector of size $L' = M * d$. This L'-dimensional supervector space is where an SVM is used to obtain a final model $\vec{w_e}$ of the target emotion e. In this case the scoring function $s'(w_e, sv(u_{test}))$ for every testing utterance u_{test} is defined as follows

$$s'(\vec{w_e}, sv(\vec{u_{test}})) = \vec{w_e} * sv(\vec{u_{test}})^T$$

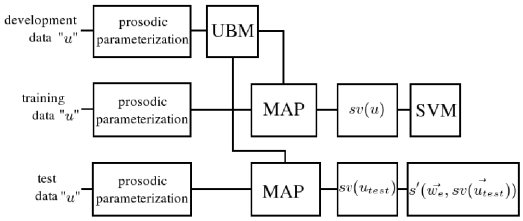

Figure 2: *Block diagram of the GMM Supervector based SVM.*

3.2. Baseline approach

The baseline system is based on a popular scheme presented in [8]. For every utterance u, the statistical distribution of the prosodic vectors is characterized by computing $n = 9$ values for each one of the prosodic streams (table 1). Thus, we obtain a $L = d * n$ fixed-length feature vector per utterance. This new derived L-dimensional feature space is where emotions are modeled by using a one-versus-all linear SVM (Figure 3. Note that this L-dimensional feature vector can be seen as the result of a kernel function $l(u)$, that maps the d-dimensional prosodic vectors of u into a L-dimensional feature space.

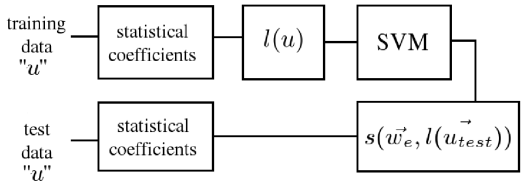

Figure 3: *Block diagram of the Baseline Classifier.*

Given an SVM model$\vec{w_e}$ of an emotion e, the scoring function $s(\vec{w}, l(\vec{u}))$ for every test utterance u_{test} is a simple dot product computed as follows:

$$s(\vec{w_e}, l(\vec{u_{test}})) = \vec{w_e} * l(\vec{u_{test}})^T$$

Table 1: *Statistical coefficients extracted for every prosodic stream in the Baseline approach.*

Coefficients
Maximum
Minimum
Mean
Standard deviation
Median
First quartile
Third quartile
Skewness
Kurtosis

On the one hand, the similarities between the proposed prosodic GMM-SVM system and the baseline are: i) Previous d-dimensional prosodic features vectors are used as inputs, ii) The modeling of their long-term statistical distribution (pdf) of the vectors in u by using linear SVMs and iii) Both cases are an attempt to characterize pdf. Nevertheless, the method used to characterize pdf's differs between both presented subsystem. As a consequence, not only performances differ, also

uncorrelated scores are generated. This fact motivates a posterior subsystem fusion in order to increase the final performance achieved. On the other hand, the baseline only uses a small set of well performing values to characterize the pdf of the vectors in every u, but probably they are not seizing the whole information embedded in it. Note for example that the baseline subsystem compute the n statistical values stream by stream, not using the correlated information among them.

3.3. Subsystem fusin

Final scores generated by the system are combinations of $s'(\vec{w_e}, sv(u_{test}))$ and $s(\vec{w_e}, sv(u_{test}))$. Combination is performed as a sum fusion preceded of a test normalization (Tnorm [ref]) stage, which fosters a similar range of the scores of both subsystems. Tnrom cohort is form by the whole set of emotions models w_e, for $e = 1...N_{emotions}$. The final combined score $S(\vec{w_e}, u_{test})$ is computed as follows

$$S(\vec{w_e}, u_{test}) = \frac{s'(\vec{w_e}, sv(u_{test})) - \mu'}{std'} + \frac{s(\vec{w_e}, sv(u_{test})) - \mu}{std}$$

Where μ' and μ are the means of the cohort scores, and std' and std the standard deviations. Referred to the Proposed and Baseline systems respectively.

4. Experiments

4.1. Databases

The proposed emotion recognition system has been tested over the English SUSAS database (Speech Under Simulated And Actual Stress). SUSAS has been employed frequently in the study of the effects of speech production and recognition, when speaking under stressed conditions [8]. This database was designed originally by John H.L. Hansen, et al. in 1998 for speech recognition under stress. All speech files from SUSAS database were sampled at 8kHz, and 16-bit integers. SUSAS Simulated subcorpora contains speech from 9 speakers and 11 speaking styles. They include 7 simulated styles ($slow$, $fast$, $soft$, $question$, $clear\ enunciation$, $angry$) and four other styles under different workload conditions ($high$, $cond70$, $cond50$, $moderate$). SUSAS Actual speech contains speech from 11 speakers, and 5 different and real stress conditions ($neutral$, $medst$, $hist$, $freefall$, $scream$). Actual and Simulated subcorpora contains 35 spoken words with 2 realisation of each, for every speaker and speaking style. The SUSAS database has been selected for the following reasons: i) presents a large set of target emotions; ii) allows comparisons with previous work in the literature; iii) speaker IDs are available; and iv) there exist simulated and actual emotional states. These two last subcorpora, namely Simulated and Actual, have characteristics different enough to consider them as different databases.

4.2. Results

Speaker inter-variability can cause that different emotions and different speakers may be located in the same region in the feature space. This drawback can be compensated by using speaker independent emotion models. To compare the performance improvement between both scenarios, we carried out speaker dependent and speaker independent experiments. Experiments are performed for both SUSAS subcorporas, Simulated and Actual. Both subcorpus have been divided in three non-overlapped sets with equivalent amount of data: training set, testing set, and a development set used for UBM training.

Any model $w_e(spk)$ or $w'_e(spk)$, for the baseline and the proposed prosodic GMM-SVM subsystems respectively, will be denoted as $w_e(spk)$ for simplicity. Performance results will be measured in terms of equal error rate (EER), which is a popular performance measure for any detection task.

4.2.1. Speaker Independent Experiments

For detection of target emotion e, every model w_e is trained using data belonging to e as the target class, and any other emotion as the non-target class. Therefore we will obtain 11 emotion models for Simulated speech and 5 models for Actual speech. In order to obtain results not affected by speaker overfitting, training, testing, and development sets, each experimental subset of SUSAS will be built with different speakers.

Table 2: *EER(%) in Speaker Independent experiments for SUSAS Simulated speech. R.I. denotes the relative improvement of $Combine$ in respect of $Baseline$.*

Emotion	Baseline	Proposed	Combined	R.I. %
angry	18.16	20.47	16.73	+7.87
clear	42.68	31.04	31.99	+25.05
cond50	40.76	39.84	38.22	+6.23
cond70	42.28	40.21	40.43	+4.37
fast	24.31	27.23	20.63	+15.13
lombard	51.24	42.06	42.55	+16.96
loud	23.03	24.57	21.03	+8.68
neutral	36.29	35.33	34.38	+5.26
question	12.44	4.38	4.38	+64.79
slow	19.60	26.10	22.46	-14.59
soft	20.65	38.19	22.26	-7.79
Avg. EER	30.13	29.94	26.82	+10.37

Table 3: *EER(%) in Speaker Independent experiments for SUSAS Actual speech.*

Emotion	Baseline	Proposed	Combined	R.I. %
neutral	35.12	34.61	33.31	+5.15
medst	40.99	42.21	41.51	-1.26
hist	36.82	38.97	35.75	+2.9
freefall	25.07	54.75	31.29	-24.81
scream	6.46	11.68	7.6	-17.64
Avg. EER	28.89	36.04	29.78	-3.08

Results in tables 2 and 3 shows better performance for Actual subcorpus than for Simulated one. This fact is probably cuased by the less number of target classes, which makes the performance of the detection of a target emotion with respect to the rest easier. Also note that the EER for similar classes such as $cond50$, $cond70$ and $lombard$ is higher than for other more differentiable emotions such as $question$ and $angry$. This emphasizes the strong dependence of the performance on the emotion set.

4.2.2. Speaker Dependent Experiments

For a speaker spk and a target emotion e, every model $w_e(spk)$ is trained using all the utterances belonging to simultaneously spk and e for the target model. Non-target model is trained in this scenario using data from all speakers and emotions except those included in the target model training set.

Table 4: *EER(%) in Speaker Dependent experiments for SUSAS Simulated speech.*

Emotion	Baseline	Proposed	Combined	R.I. %
angry	11.07	12.00	9.04	+18.33
clear	37.51	26.31	26.34	+29.77
cond50	37.40	33.61	32.38	+13.42
cond70	37.17	33.52	33.14	+10.84
fast	20.18	19.71	15.62	+22.59
lombard	31.14	29.02	26.63	+14.48
loud	15.56	11.27	10.17	+34.64
neutral	32.22	27.31	26.04	+19.18
question	5.80	3.19	1.98	+65.86
slow	16.66	15.08	13.17	+20.94
soft	10.13	15.67	10.18	-0.49
Avg. EER	23.16	19.70	18.60	+19.68

Table 5: *EER(%) in Speaker Dependent experiments for SUSAS Actual speech.*

Emotion	Baseline	Proposed	Combined	R.I. %
neutral	18.23	17.21	15.23	+16.45
medst	27.06	24.29	22.79	+15.77
hist	23.35	21.53	19.85	+14.98
freefall	25.40	19.27	20.97	+17.44
scream	8.31	5.72	5.72	+31.16
Avg. EER	20.47	17.60	16.91	+17.39

Results in tables 4 and 5 shows that by combining individual classifiers in a speaker dependent framework, we can achieve better performance than for any of them separately. Relative improvements of the combined approach respect to the baseline are about 17.4% or 19.7% in Actual and Simulated speech respectively. Table 6 also shows that class overlapping is remarkable reduced between speaker dependent and independent schemes. Note that the Combined system achieves a relative improvement about 30.64% when it is evaluated in Actual subcorpus. Relative improvement is about 43.21% for Simulated subcorpus.

5. Conclusions

This work introduces a novel approach for emotion recognition using prosodic features. The porposed approaches models the statistical distribution of short-term pitch, energy and their velocities by a GMM, and the a SVM classification of in the mean-supervector space of the models gives the final score for detection. We compare this prosodic GMM-SVM system with a baseline implementing a popular approach also at the prosodic level. Moreover, we explore a combination (fusion) approach with a baseline system, which further increases performance. The task is presented as a verification or detection problem measured in terms of EER. The experimental set-up is based on two subcorpus of the SUSAS database, as well as in two different experimental frameworks: speaker-independent and speaker-dependent. According to results we conclude that the proposed approach achieved equal or better results than the baseline. Remarkably enough, the fusion of both approaches in a speaker-dependent framework yields performance improvements by a factor of 17.4% or 19.7% respectively for Actual and Simulated subcorpus. We also conclude that by removing speaker inter-variability the system performance significantly improves. The relative improvement is about 30.64% when it is evaluated in Actual subcorpus and about 43.21% for Simulated subcorpus.

Table 6: *Comparation between speaker independent and speaker dependent experiments*

Subcorpus	Approach	Spk. Ind.	Spk. Dep.	R.I.%
Actual	Baseline	30.13	23.16	+23.13
	Proposed	29.94	19.70	+34.20
	Combined	26.82	18.60	+30.64
Simulated	Baseline	28.89	20.47	+29.14
	Proposed	36.04	17.0	+52.83
	Combined	29.78	16.91	+43.21

The use of new improved configurations for pitch continuous estimation will be addressed in future work as well as the combination of prosodic and acoustic level of features.

6. References

[1] Rosalind W. Picard, *Affective Computing*, The MIT Press, September 1997.

[2] L.C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information", Sep 1997, vol. 1, pp. 397–401 vol.1.

[3] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge", 2009.

[4] Zhihong Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[5] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-case database in spanish", in *Proceedings of Interspeech 2008*, September 2008, pp. 1493–1496.

[6] J.H.L. Hansen and S.E. Bou-Ghazale, "Getting started with susas: a speech under simulated and actual stress database", in *EUROSPEECH-1997*, 1997, pp. 1743–1746.

[7] J.H.L. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition", in *Speaker Classification (1)*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 108–137, Springer.

[8] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee, "Emotion recognition by speech signals", in *EUROSPEECH-2003*, 2003, pp. 125–128.

[9] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.04) [computer program]", Ap 2009, http://www.praat.org/.

[10] Hao Hu, Ming-Xing Xu, and Wei Wu, "Gmm supervector based svm with spectral features for speech emotion recognition", in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–413–IV–416.

Anchor Model Fusion for Emotion Recognition in Speech

Carlos Ortego-Resa, Ignacio Lopez-Moreno
, Joaquin Gonzalez-Rodriguez, and Daniel Ramos

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
carlos.ortego@estudiantes.uam.es,
http://atvs.ii.uam.es/

Abstract. **Key words:** emotion recognition, anchor models, backend, prosody, GMM supervectors, SVM.

1 Introduction

Automatic emotion recognition in speech is ganing a strong support in the scientific community due to its applications to human-machine interaction industry [1]. As a result new methodologies focused on a wide range of information sources and classification schemes have emerged. This fact motivates the use of fusion schemes that seizes uncorrelated information of each scheme.

It is common for this task to be stated as a multiclass classification problem. However, emotion recognition can also be headed as a verification or detection problem. In such case, given an utterance x and a target emotional state e, from a N_{fe} emotions set, the objective is to determine whether the dominant emotion that affect the speaker in the utterance is e (*target* class) or not (*non-target* class). In such squeme any model m_e, and utterance x, can compute a similarity score denoted as $s_{\mathbf{x},m_e}$. Classification is performance by comparing $s_{\mathbf{x},m_e}$ to a given threshold. In this work models $M = [m_j]$, $j \in \{1, .., N_{fe}\}$ are denoted as front-end models in oposition to back-end models which are trained in advance using scores, such as $s_{\mathbf{x},m_j}$, as feature vectors.

Consider that limits among emotions may not be clear and often overlaped, moreover when different databases and different target emotions are taken into account. This fact leads, for models of different emotions, to characteristicaly rate when they are compare to utterances of any emotion e. And not only when they are compare with the target model. We expect for models in M to offer additional information that back-end emotion models can learn.

This work propuses a novel back-end approach that combines outputs from N_{sys} different classification schemes. It is based on *anchor models* [2] and supports the final decission not only on the target emotion model but also on the relationship among all the avaible models in M.

In order to show the viability of this novelty technique in various embiroments, three emotional labeled corporas have been used: *Ahumada III* [3]., *SUSAS Simulated* and *SUSAS Actual* [4]. AMF have been used to combine scores from

two prosidic emotion recognition systems denoted as GMM-SVM and statistics-SVM. Performance results will be measured in terms of equal error rate (EER), average EER and relative improvement in the EER, which are popular performance measures for any detection task.

This work is organised as follows. The role of anchor models described in Section 2. In Section 3, the proposed AMF system is described in detail. Section 4 describes front-end systems implemented as well as the prosodic parametrizatin. The experimental work which shows the adequacy of the approach is shown in 5.2. Finally, conclusions are drawn in Section 6.

2 Anchor models

Given a speech utterance x from a unknown spoken emotion, and a front-end emotion recogtion system that models N_{fe} target emotions $M = [m_j]$, $j \in \{1, .., N_{fe}\}$. A similarity score $s_{\mathbf{x},m_j}$, can be obtain as a result of comparing x against any emotion model m_j.

Consider that m_j is replaced by all the models in M. In this case, for every utterance x we obtain a N_{fe} dimensional vector $\bar{S}_{\mathbf{x},\mathbf{M}}$ that stacks all posible values of $s_{m_j,\mathbf{x}}, j \in \{1, .., N_{fe}\}$.

$$\bar{S}_{\mathbf{x},\mathbf{M}} = [s_{\mathbf{x},m_1} \cdots s_{\mathbf{x},m_N}] \tag{1}$$

This scheme defines a derived similarity feature space known as *anchor model* space in which every utterance x can be projected. The anchor model projection allows for back-end data driven classifiers, to train in advance new emotion models $M' = [m'_j]$, $j \in \{1, .., N_{be}\}$, by learning the realative behavior of the speech utterance x with respect to M. This relative behaviour is shown in figure 1 where utterances from four emotions (*angry*, *question*, *neutral*, *extressed*,) are compared with two different cohorts M of anchor models.

Notice that the N_{fe} front-end models in M do not need to match with the N_{be} target models in the back-end stage, denoted as M'. However, feature vectors from the target emotions models in the back-end stage M' require to behave distinguishably with respect to models in M.

3 Anchor Model Fusion (AMF) back-end

AMF is a data-driven approach that have shown an excelent performance when it is applied in language recongnition phone-SVM models [5]. In AMF, the cohort of models M is built by including all the available models from the N_{sys} emotion recognition systems in the front-end. Resulting AMF similaritie vector of the utterance x, denoted as S_{AM}, stacks the N_{sys} values of $S^j_{x,M}$ over all emotion recognition system j in the front-end.

$$\bar{S_{AM}}(x, M) = \left[\bar{S}^1_{x,m}, \cdots \bar{S}^{N_{sys}}_{x,m}\right] \tag{2}$$

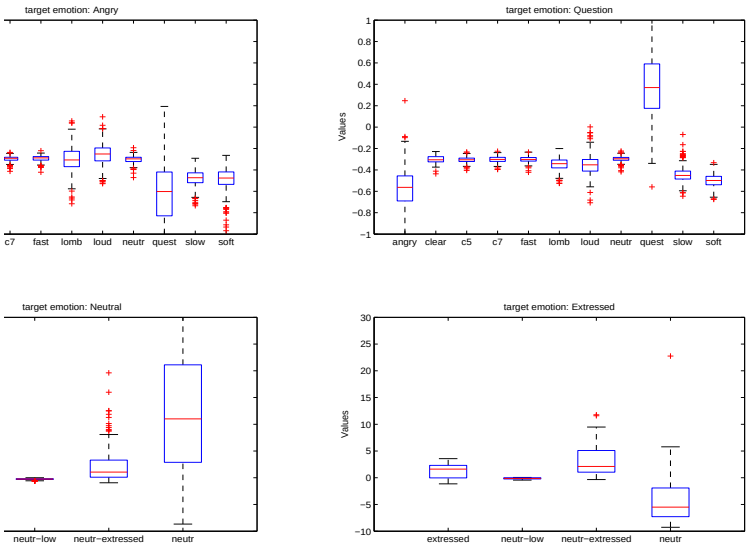

Fig. 1. *Up. Relative range of* angry *(left) and* question *(right) utterances over the a set M form by the emotion models in SUSAS Simulated speech. Down. Relative range of* neutral *(left) and* extressed *(right) utterances over the a set M form by the emotion models in Ahumada III*

.

Fig. 2 ilutrate the process in which $S_{AM}(x, M)$ is obtained by projecting x into the AMF space defined by M.

Hence, the number of dimensions of AMF space is $d = \sum_{j=1}^{N_{sys}} N_j$, where N_j is the number of models in the front-end system j. At this point, the objective is to boost the probability of finding a characteristic behavior of the speech pattern in the anchor model space, by increasing d. This objective can be achieved by different and complementary approaches: i) Including in M front-end models of the back-end target emotions ($M' \in M$). ii) Including in M models from different databases, and techniques, such as Gaussian Mixture Models (GMM), SVM, n-grams, etc [6]. iii) Including in M hierarchy emotion models. The following example ilustrate this situation. Consider that our goal is to separate between $extressed$ and $non-extressed$ speakers, by uncluding in M models of emotions such as *happy*, *anxious* or *angry* back-end results will be supported by the behavior of $extressed$ and $non-extressed$ utterances over these hierarchily lower, emotion models.

Once every training and testing utterance is projected over the AMF space, any classifier can be used for training any back-end emotion in M'. In this

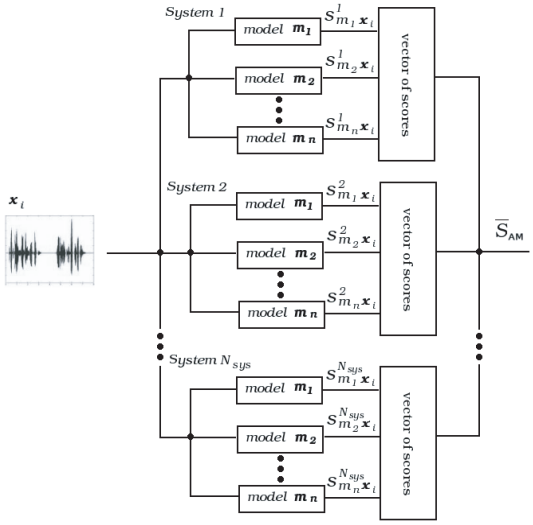

Fig. 2. *Diagram of generation of features in the AMF space.* $\bar{S}_{AM}(x, M)$ *stacks the similarities of* $\boldsymbol{x}_i$ *over the set of models* m^l_j*, for language* j *and subsystem* l

.

work, SVM were applied due to its robustness while the dimension of the AMF increases.

4 Emotion recognition systems fron-end

This section details the prosodic parametres extracted from the audio signal, and used as input vectors for both front-end systems implemented. Subsections 4.2 and 4.3 descrives in more detail their implementation.

4.1 Prosodic features for emotion recogniton

Prosodic features are often considered as input signals for emotion recognition systems due to their relation with the emotional state information [4]. In this work prosodic features consist of a set of $d = 4$ dimensional vectors with the sort-term coefficients of energy, the logarithm of the pitch and their velocity coefficients, also known as Δ features. These coefficients are extracted only from voiced segments with an energy value higher than the 90% of the dinamic range. Mean normalization have been used for energy and Δ-energy coefficients. Pitch and energy have been computed by using Praat [7].

4.2 prosodic GMM-SVM

Previous works have shown the excelet performance of SVM-GMM supervectors in the tasks of language and speaker recognition, while the application of this technique to the prosodic level of the speech were firstly introduced in [8].

This technique can be seen as a secondary parametrization capable to summarize the distribution of the feature vectors in x, into a single high-dimensionality vector. This high-dimensionality vector is known as a GMM supervector. In order to build the GMM supervector, first the prosodic vectors of x are used to train a M-mixtures GMM model λ_x, as a Maximun A Posteriori (AMP) adaptaion of means from a general GMM model λ_{UBM}. The GMM supervector of the utterance x is the concatenation of the M vectors of means in λ_x.

GMM supervector are often considered as kernel functions $\mu(x)$ that maps prosodic features from dimension of d into a high-dimensional feature space of size $L' = M * d$. Once every utterance is mapped into this L'-dimensional supervector space, linear SVM models are used to train the front-end emotion models. Therefore, any m_j is a L'-dimensional vector that represent an hiperplane that optimally, separate supervectors of utterances form the target emotion j with respect to supervectors from other emotions.

4.3 prosodic statistics-SVM

This scheme is based on a previous work presented in [9]. It consist on a statistical analisys of each prosodic coefficient followed by a SVM. The distridution of the prosodic values is charaztericed by computing $n = 9$ statistical coefficients per feature (table 1). Once every utterance is mapped into this derived feature space of dimension $L = d * n$, front-end emotions models are obtained as linear one-vs-all SVM models.

Table 1. *Statistical coefficients extracted for every prosodic stream in the statistics-SVM approach.*

Coefficients
Maximum
Minimum
Mean
Standard deviation
Median
First quartile
Third quartile
Skewness
Kurtosis

It is common for systems presented in sections 4.3 and 4.2 to generate scores in different ranks. This fact motivates the use of a posterior score normalization

technique before they are used to built AMF feature vectors. Test normalization (Tnorm [ref]) have been used for this purpose. Tnorm estimate the scores distribution for every testing utterance x_t by comparing x_t over a cohort of models. The values of mean and variance of this distribution are then used to normalise the similarity scores of x_t over any model m_j. In this work M have also been used as Tnorm cohort.

5 Experiments

5.1 Databases

The proposed emotion recognition system has been tested over Ahumada III and SUSAS (Speech Under Simulated And Actual Stress) databases. Ahumada III is form by real forensics cases recorded by the spanish police forces (*Guardia Civil*). It includes speech from 69 speakers and 4 emotional states (*neutral*, *neutral-low*, *neutral-extressed*, *extressed*) with 150 seconds training utterances while testing utterances are 10 and 5 seconds lenght. SUSAS database is divided in two subcorpora from simulated and real spoken emotions. SUSAS Simulated subcorpora contains speech from 9 speakers and 11 speaking styles. They include 7 simulated styles (*slow*, $fast$, $soft$, *question*, *clear enunciation*, *angry*) and four other styles under different workload conditions (*high*, *cond*70, *cond*50, *moderate*). SUSAS Actual speech contains speech from 11 speakers, and 5 different and real stress conditions (*neutral*, *medst*, *hist*, $freefall$, *scream*). Actual and Simulated subcorpora contains 35 spoken words with 2 realisation of each, for every speaker and speaking style.

5.2 Results

Experiments were carry out over corpora presented in section 5.1 and systems presented in sections 4.2 and 4.3.

The GMM-SVM front-end system requires a set of development data for building the model λ_{UBM}. Therefore every database were splited in two different and non overlaped sets. The first one have been used for training a M=256 mixtures GMM model (λ_{UBM}). For this purpose we used Expecteation Maximization (EM) algorithm. The second set were used for implemeneting two stages of boot straping. A first stage is used for training and testing front-end models, while back-end models are trained and tested during the second stage. These two stages of boot straping repectively used a 90% and 10% of the available data for training and testing purposes.

AMF cohort M is form with models from all databases and systems. Therefore for each one of both front-end system we obatined 4 models from Ahumada corpus, 11 models from SUSAS Simulated corpus and 5 models from SUSAS Actual corpus. M includes models for both systems as well as their sum fusion, this scheme leads to a AMF space of $(4 + 11 + 5) \times 3 = 60$ dimensions.

In order to compare AMF with a *baseline* fusion technique we performed a standard sum fusion between the scores of GMM-SVM and statistics-SVM

systems. Notice that sum fusion outcomes the results obtained from any of both system individaualy.

Table 2. *Comparation between AMF and sum fusion both implemented emotion recognition systems. Results in terms of EER(%) and relative improvement (R.I.) for* SUSAS Simulated*,* SUSAS Simulated *and* Ahumada III

SUSAS Simulated			
Emotion	Baseline	AMF	R.I. %
angry	22.93	32.76	42.87
clear	42.91	41.89	-2.38
cond50	41.01	33.57	-18.14
cond70	48.3	30.55	-36.75
fast	30.21	16.81	-44.36
lombard	34.85	38.65	10.9
loud	27.65	13.2	-52.26
neutral	40.53	35.31	-12.88
question	3.86	3.52	-8.81
slow	26.75	20.35	-23.93
soft	22.07	22.54	2.13
Avg. EER	31.01	26.29	-15.22

SUSAS Actual			
Emotion	Baseline	AMF	R.I. %
neutral	36.54	35.26	-3.5
medst	46.95	50.08	6.67
hist	42.57	39.14	-8.06
freefall	25.86	24.66	-4.64
scream	11.15	14.6	30.94
Avg. EER	32.61	32.75	0.43

AhumadaIII			
Emotion	Baseline	AMF	R.I. %
neutral-low	50.21	30.02	-40.21
neutral	33.77	33.92	0.44
neutral-extressed	38.12	33.22	-12.85
extressed	28.69	25.7	-10.42
Avg. EER	37.7	30.72	-18.51

Obtained results over *Ahumada III* and *SUSAS Simulated* (table 2) shows an average improvement larger than a 15%. Remarkable good results are obtained for *neutral-low*, *loud* and *fast* emotion models while for models *scream* and *angry* a significant loss of performance is obtained, probably due to non modeled variablity factors such as the speaker identity.

6 Conclusions

This work introduces a novel approach for combining outputs from N_{sys} emotion recognition systems in a robust way. The approach is based on the anchor model space which defines a derived feature space where new back-end models can be trained in advance. When anchor models are used for fusing a set of front-end systems, similarities over a all their models are used as features. Therefore back-end emotion models m' are supported over the set of front-end models M trained with different emotions, databases, recording conditions, etc. In this work the proposed AMF approach have been used for fusing two different prosodic emotion recognition systems as well as a third one obtained as the result of the sum fusion of both systems. Thus M have been built with 3 systems and 20 front-end models which leads to a 60-dimensions AMF space. Experiments have been carry out over three corpora (*Ahumada III*, *SUSAS Simulated* and *SUSAS Actual*) with simutated and real emotions, different languges and recordin conditions. Resuts are compared with the sum fusion of both front-end systems. They show a performance improvement larger than the 15% for *Ahumada III* and *SUSAS Simulated* corpora. Future work will explore on the optimal selection of models in M, normalization techniques of the AMF space vectors and new classification methods such as Linear Discriminant Analysis.

References

1. Picard, R.W.: Affective Computing. The MIT Press (September 1997)
2. Collet, M., Mami, Y., Charlet, D., Bimbot, F.: Probabilistic anchor models approach for speaker verification. (2005) 2005–2008
3. Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Lucena-Molina, J.J.: Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-case database in spanish. In: Proceedings of Interspeech 2008. (September 2008) 1493–1496
4. Hansen, J., Patil, S.: Speech under stress: Analysis, modeling and recognition. In: Speaker Classification (1). Volume 4343 of Lecture Notes in Computer Science., Springer (2007) 108–137
5. Lopez-Moreno, I., Ramos, D., Gonzalez-Rodriguez, J., Toledano, D.T.: Anchor-model fusion for language recognition. In: Proceedings of Interspeech 2008. (September 2008)
6. Benesty, J., Sondhi, M.M., Huang, Y.E.: Springer Handbook of Speech Processing. Part G. Springer (2008)
7. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (version 5.1.04) [computer program] (Ap 2009) http://www.praat.org/.
8. Hu, H., Xu, M.X., Wu, W.: Gmm supervector based svm with spectral features for speech emotion recognition. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Volume 4. (2007) IV–413–IV–416
9. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion recognition by speech signals. In: EUROSPEECH-2003. (2003) 125–128

B
Presupuesto

1) Ejecución Material

▪ Compra de ordenador personal (Software incluido)	2.000 €
▪ Alquiler de impresora láser durante 10 meses	200 €
▪ Material de oficina	150 €
▪ Total de ejecución material	2.350 €

2) Gastos generales

▪ 16 % sobre Ejecución Material	376 €

3) Beneficio Industrial

▪ 6 % sobre Ejecución Material	141 €

4) Honorarios Proyecto

▪ 1000 horas a 15 €/ hora	15000 €

5) Material fungible

▪ Gastos de impresión	200 €
▪ Encuadernación	100 €

6) Subtotal del presupuesto

▪ Subtotal Presupuesto	18.167 €

7) I.V.A. aplicable

▪ 16 % Subtotal Presupuesto	2.906.72 €

8) Total presupuesto

▪ Total Presupuesto	21073.72 €

Madrid, Julio 2009
El Ingeniero Jefe de Proyecto

Fdo.: Carlos Ortego Resa
Ingeniero Superior de Telecomunicación

C

Pliego de condiciones

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, *DETECCIÓN DE EMOCIONES EN VOZ ESPONTÁNEA*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales.

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirviá de base para la contratación, a las modificaciones autorizadas por la superioridad o a

las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contratać anteriormente llamado "Presupuesto de Ejecución Material"que hoy designa otro concepto.

Condiciones particulares.

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.