UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

MÁQUINAS DE VECTORES SOPORTE (SVM) PARA RECONOCIMIENTO DE LOCUTOR E IDIOMA

-RESUMEN DEL PROYECTO FIN DE CARRERA-

XXVIII Convocatoria premios "Ingenieros de Telecomunicación"

Ismael Mateos García
Febrero de 2008

1. Descripción del proyecto

1.1 Origen

Los rápidos avances llevados a cabo en el campo de las redes de comunicación y la movilidad han propiciado la aparición de un nuevo conjunto de *tele-aplicaciones*. Dentro de este nuevo conjunto están englobadas todas aquellas aplicaciones que permiten una comunicación remota entre el usuario y cualquier tipo de sistema. La banca telefónica o la venta de entradas *on-line*, son sólo algunos ejemplos.

Todo este tipo de aplicaciones requiere una autenticación por parte del usuario. Tradicionalmente, se empleaban esquemas de identificación clásicos, los cuales hacían uso de claves secretas, códigos o llaves. En la actualidad, los sistemas basados en reconocimiento biométrico se presentan como una buena alternativa a los métodos clásicos.

Entre las principales ventajas de este tipo de sistemas podemos destacar, su bajo coste de mantenimiento, el alto nivel de seguridad que ofrecen y la comodidad para el usuario. Mientras que las claves de los sistemas clásicos eran fácilmente olvidables los rasgos biométricos, como por ejemplo la voz, huella dactilar, etc. son características que siempre porta consigo el individuo.

Para que una característica o comportamiento sea considerado rasgo biométrico deberá cumplir una serie de propiedades. La voz es un rasgo biométrico que además de cumplir estos requisitos cuenta con muchas ventajas, entre las que podemos destacar su fácil adquisición y transmisión. Puede ser adquirida y transmitida de una manera muy sencilla, sin métodos invasivos ni dispositivos especializados, basta con un simple micrófono y un canal de transmisión convencional (telefonía fija, telefonía móvil, redes IP, etc.). Estas características hacen de la voz un rasgo biométrico ideal para aplicaciones remotas.

Otra de las características de la señal de voz es la gran cantidad de información que contiene: identidad del locutor, idioma, edad, estado de ánimo, nivel de educación, etc. Los sistemas de reconocimiento biométrico harán uso de la información sobre la identidad del locutor para identificar a los usuarios. Por otra parte, la información acerca del idioma del hablante será importante para aplicaciones orientadas a la seguridad, información, adaptación a usuarios, etc.

El reconocimiento automático del idioma comparte muchas técnicas con el reconocimiento de locutor, por tanto ambos problemas podrán ser abordados de un modo similar.

1.2 Objetivo y enfoque

El presente proyecto viene motivado por la creciente demanda de aplicaciones remotas que hace cada vez más evidente la necesidad de mejorar e investigar en el campo de los sistemas de reconocimiento automático. Por otro lado, la proliferación de servicios telefónicos multilingües requiere de sistemas de detección de idioma en voz espontánea, dichos sistemas ayudan a tratar convenientemente cualquier llamada en un tiempo reducido. Esta motivación lleva a que en el proyecto se aborden dos problemas relacionados con la señal de voz: por un lado el reconocimiento biométrico basado en dicha señal y por otro el reconocimiento del idioma del hablante.

El objetivo del proyecto será presentar e implementar nuevos métodos de reconocimiento de locutor e idioma, además de realizar un repaso por el estado del arte de las técnicas existentes. El trabajo se centra en el uso de Máquinas de Vectores Soporte (SVM), una técnica de modelado consolidada en el campo del reconocimiento de patrones, cuya eficiencia ha sido ampliamente demostrada en los últimos años [Campbell *et al.,* 2006b]. Se realizará un examen completo de estos sistemas, desde las formas de extraer las características de la señal de voz, el conjunto de datos de entrenamiento, influencia de distintas variables en los modelos entrenados, fusión de sistemas, etc.

El proyecto presenta un extenso estudio experimental, cuyos resultados avalan que tanto las propuestas originales como las técnicas empleadas son eficaces en las tareas llevadas a cabo. Los experimentos se realizan siguiendo el protocolo de evaluaciones NIST *(National Institute of Standards and Technology)*. Estas evaluaciones tienen un carácter abierto, en ellas participan grupos de investigación de todo el mundo, constituyendo de este modo un foro científico y tecnológico que ha impulsado el desarrollo de los sistemas de reconocimiento basados en voz en la última década. Por último, se presentan las conclusiones y proponen las líneas de trabajo futuras.

1.3 Desarrollo

El proyecto comienza realizando una introducción a los sistemas de reconocimiento automático, sistemas donde se incluye el reconocimiento de locutor e idioma. Se explican los distintos modos de funcionamiento de este tipo de sistemas y se realiza una breve descripción de los rasgos biométricos más habituales, destacando las principales ventajas de la voz.

Las secciones siguientes del proyecto realizan un recorrido por el estado del estado del arte del reconocimiento biométrico basado en la señal de voz y del reconocimiento automático del idioma. A continuación se detalla el sistema implementado en el proyecto, características empleadas y fundamento teórico del SVM. Además se presenta el protocolo empleado para la evaluación objetiva de los sistemas.

1.3.1 Extracción de características en locutor e idioma

La extracción de características es el paso previo a cualquier sistema de reconocimiento automático. En primer lugar se capta la señal sobre la que se desea trabajar mediante un sensor, en este caso será un sensor apto para la señal de voz (micrófono, teléfono, etc.).

La extracción de parámetros está basada habitualmente en el análisis a corto plazo de la señal de voz, para ello una de las técnicas más habituales en reconocimiento automático de locutor es MFCC *(Mel-Frequency Cepstral Coefficients)*, en reconocimiento de idioma haremos uso de esta técnica y de otra conocida como SDC *(Shifted Delta Cepstral)*, las cuales pasaremos a detallar más adelante.

Sobre estas técnicas pueden llevarse a cabo mejoras con el fin de paliar las distorsiones sufridas por la señal de voz y mejorar el rendimiento de los sistemas, estas mejoras son conocidas como compensaciones de canal y en el trabajo se utilizarán cuatro de las más importantes: CMN *(Central Mean Normalization)* [Furui, 1981], *Feature Mapping* [Reynolds, 2003], *Feature Warping* [Pelecanos y Sridharan, 2001] y *RASTA filtering* [Hermansky y Morgan, 1994].

MFCC *(Mel-Frequency Cepstral Coefficients)*

Los coeficientes MFCC se extraen a partir de la representación de la señal de voz en el dominio espectral [Deller *et al.*, 1999]. Diversas investigaciones llevadas a cabo hasta la fecha, han demostrado que los coeficientes obtenidos del dominio espectral representan más fielmente las características de la voz que los obtenidos del dominio temporal. Esta peculiaridad es debida a que las personas utilizan este mismo dominio para distinguir sonidos, por tanto, cabe esperar que un sistema que trabaje con esas características se acerque más al comportamiento humano.

La Figura 1 representa mediante un diagrama de bloques el proceso aplicado a la señal de voz para obtener los coeficientes MFCC. A parte de estos coeficientes se suelen utilizar otros conocidos como deltas o coeficientes de primera y segunda derivada. Estos coeficientes tratan de representar la información de coarticulación entre fonemas, para ello miden velocidades y aceleraciones alrededor del instante de tiempo dado. La Figura 2 muestra un ejemplo del cálculo de los coeficientes delta.

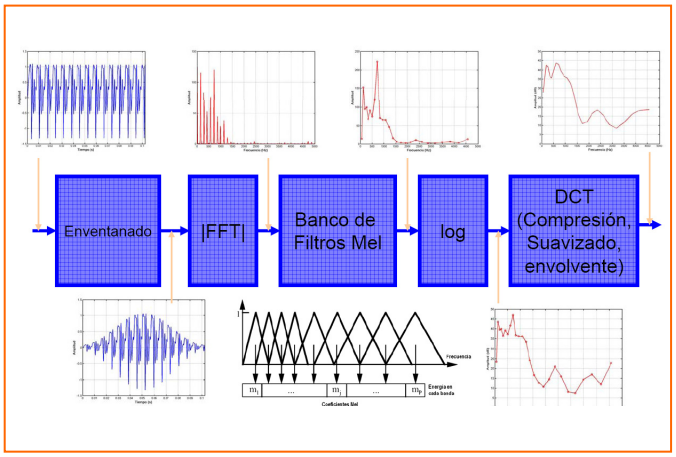

Figura 1. Extracción de coeficientes MFCC

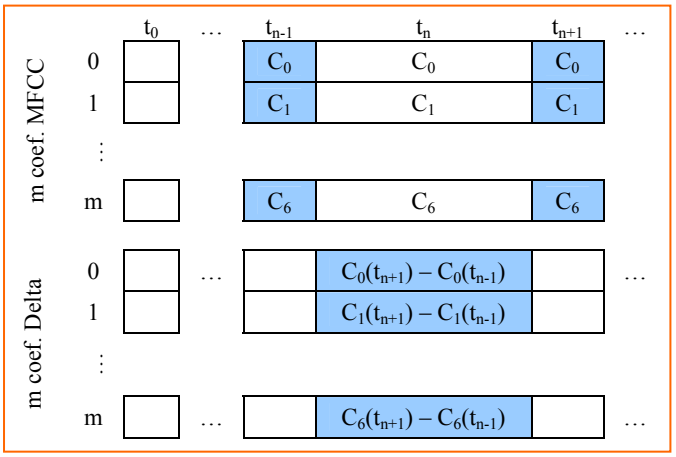

Figura 2. Ejemplo cálculo coeficientes Delta

SDC *(Shifted Delta Cepstral)*

La parametrización SDC [Torres-Carrasquillo, 2002] puede tratarse como una parametrización derivada de la MFCC. La información de cada vector de parámetros es función de los vectores de parámetros adyacentes, al igual que sucedía con los coeficientes delta pero se introducen mejoras.

Los coeficientes SDC vienen especificados por cuatro parámetros, N-d-P-k, donde: **N** es el número de coeficientes cepstrales, **d** representa el desplazamiento en tiempo para el cálculo de las deltas (hacia delante y hacia atrás), **P** es el desplazamiento entre bloques consecutivo, por último, **k** es el número bloques que serán concatenados para formar el vector final. Para cada instante *t* los coeficientes SDC para cada bloque C_n se calculan siguiendo:

$$\Delta C_n(t,i) = C_n(t+iP+d) - C_n(t+iP-d)$$

$$con: \quad n = 0,....,N-1 \quad y \quad i = 0,....,k-1$$

La Figura 3 muestra un pequeño ejemplo del cálculo de la parametrización SDC a partir de una MFCC de 3 coeficientes, en este caso la configuración elegida es 3-2-1-3.

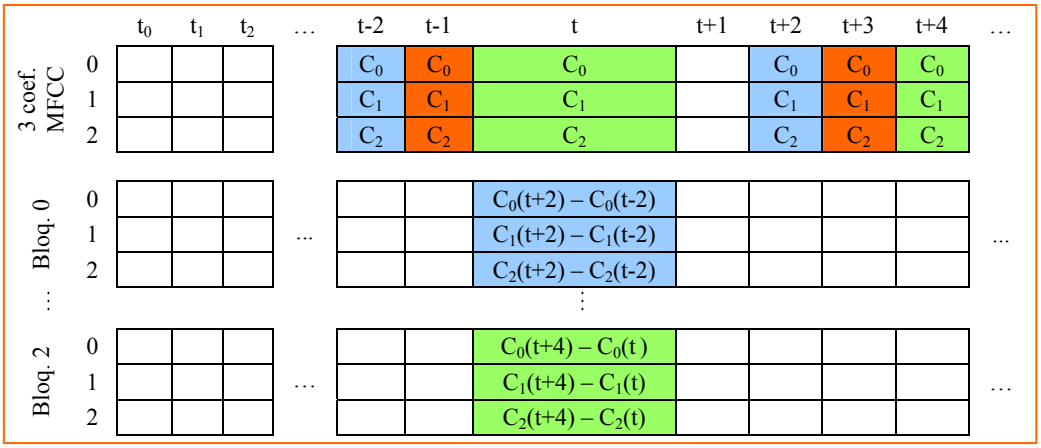

Figura 3. Ejemplificación del cálculo de parámetros SDC 3-2-1-3

1.3.2 Máquinas de Vectores Soporte (SVM)

Las máquinas de vectores soporte son básicamente un algoritmo de clasificación de patrones binario, cuyo objetivo es asignar cada patrón a una clase [Campbell *et al.*, 2006b]. Por ejemplo, si tenemos dos conjuntos de elementos, uno de ellos compuesto por ovejas blancas y otro por ovejas negras, el algoritmo tratará de diferenciar estas ovejas en función de su color (clase).

Comenzaremos la explicación del fundamento teórico del algoritmo con el caso más simple, vectores linealmente separables. Los datos con los que entrenaremos el sistema serán una serie de vectores etiquetados, de la forma: $\{\vec{x}_i, y_i\} \quad i = 1,....,l$

Donde: $\vec{x}_i \in R^d$ es el vector de observaciones en un espacio de dimensión d

$\vec{y}_i \in \{-1, 1\}$ es la etiqueta de la clase a la que pertenece cada vector

El problema consistirá en asignar cada vector de test a su clase correspondiente, 1 ó -1, para ello se construirá un hiperplano de separación que divida el espacio R^d en dos regiones. A la hora de formular el problema formalmente supondremos que todos los datos de entrenamiento cumplen una de las siguientes restricciones:

$$\vec{x}_i \cdot \vec{w} + b \geq +1 \quad si\ y_i = +1$$
$$\vec{x}_i \cdot \vec{w} + b \leq -1 \quad si\ y_i = -1$$

Combinando las dos restricciones en una: $y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \forall i$

Donde: $\vec{w}$ es un vector normal al hiperplano de separación y b es una constante

El hiperplano de separación, representado por $\vec{w}$, buscado por el algoritmo es aquel que maximiza la distancia, *m*, entre las dos clases del problema (véanse Figuras 4 y 5).

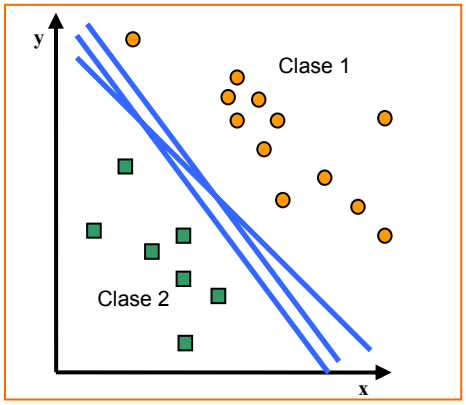

Figura 4. Posibles hiperplanos de separación

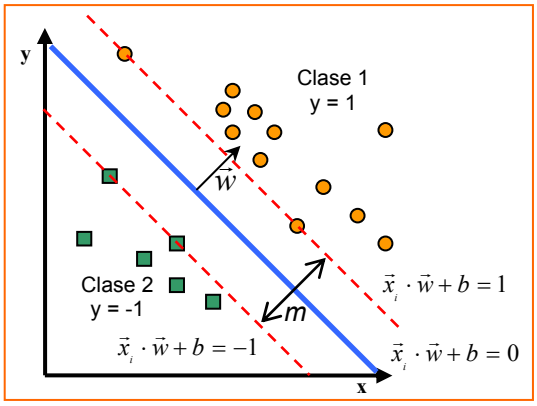

Figura 5. Hiperplano de separación óptimo

El problema de la búsqueda del mejor hiperplano de separación se traduce en un problema de optimización, existen varios métodos para su resolución, siendo uno de los más conocidos el paso a la formulación de Lagrange.

Debido a las distorsiones introducidas por el canal de transmisión, el ruido y otros efectos no deseados, puede producirse cierto solapamiento entre ambas clases. La forma en la que el algoritmo tiene en cuenta este solapamiento es a través de una función de pérdidas que intentará minimizar a la vez que se maximiza el margen.

La nueva propuesta que se introduce en el proyecto consiste en variar la forma habitual en la que se entrena el hiperplano de separación en el SVM. La mayor parte de los sistemas existentes están basados en lo explicado hasta el momento y conocido como *clasificación*, la nueva aproximación empleará la *regresión* para el entrenamiento del hiperplano [Lopez-Moreno *et al.,* 2007]. Gracias a esta técnica el sistema final será más robusto a muestras espurias y vectores con distorsiones empleados en la construcción de $\vec{w}$, pudiendo ajustarse a través de un parámetro, ε.

El objetivo de la regresión es la estimación de una función, en lugar de predecir una etiqueta $y_i = \{\pm 1\}$ como era el caso de la clasificación. Para el sistema basado en regresión, y_i será considerado como una función dependiente de $\vec{x}_i$, esta función podrá tomar cualquier valor real, a diferencia de lo que sucedía en clasificación donde sólo podía tomar los valores correspondientes a las etiquetas de las clases.

Una de las principales diferencias entre la clasificación y la regresión es la función de pérdidas, siendo la penalización más estricta en el caso del entrenamiento basado en regresión que en el basado en clasificación. La sección 6 del proyecto, *SVMs para el reconocimiento de locutor e idioma,* profundiza en una mayor medida en estos aspectos.

En ambos tipos de entrenamiento el caso de vectores no linealmente separables es afrontado de la misma manera, se aumenta la dimensionalidad del problema (teoría de la dimensión VC). Para tal fin se empleará una expansión polinómica de grado 3, seguida del kernel GLDS *(Generalized Linear Discriminative Sequence)* [Campbell, 2002].

1.3.3 Protocolos, bases de datos y presentación de resultados

El protocolo empleado para evaluar los sistemas será el definido por NIST [NIST] en sus evaluaciones más recientes en el campo de reconocimiento de locutor, NIST SRE 2006, y en el de reconocimiento de idioma, NIST LRE 2005.

La tarea fundamental de las evaluaciones NIST SRE *(Speaker Recognition Evaluation)* [NIST SRE] consiste en la verificación o detección de un determinado individuo en una grabación de prueba. Para ello, se dispone de una cantidad de datos de entrenamiento y test. La combinación entre una determinada cantidad de habla de entrenamiento y otra de prueba se denomina *condición*. El sistema implementado se probará con la condición obligatoria de todos los años, *1conv-1conv* (una conversación de habla de entrenamiento y otra de prueba de 2,5 minutos de duración).

Las evaluaciones NIST LRE *(Language Recognition Evaluation)* [NIST LRE] constan de un número de idiomas. Al igual que sucede en el campo de locutor la evaluación está compuesta por diferentes pruebas de distinta dificultad. Estas pruebas irán orientadas detectar la presencia de un idioma en la grabación de test, cuya duración es de 3, 10 ó 30 segundos. Los segmentos de 30s serán los usados en los experimentos.

Las bases de datos empleadas en reconocimiento de locutor para el entrenamiento de los modelos, normalizaciones, etc. serán NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard-I y Switchboard-II. Como datos de desarrollo para el sistema de reconocimiento de idioma, se han utilizado datos de pasadas evaluaciones NIST de idioma (NIST LRE 1996 y 2003), además de la base de datos Callfriend, disponible en [LDC].

La forma de presentar los resultados será la habitual en este tipo de evaluaciones, de manera gráfica mediante las curvas DET (*Detection Error Tradeoff*) [Martin *et al.,* 1997] y a través del EER *(Equal Error Rate)* y DCF *(Detection Cost Function)*. La Figura 6 muestra la comparativa de dos sistemas a través de su curva DET y EER, a medida que la curva está más cerca del origen, menor será la tasa de error. En el ejemplo el sistema 2 presenta un comportamiento superior al sistema 1.

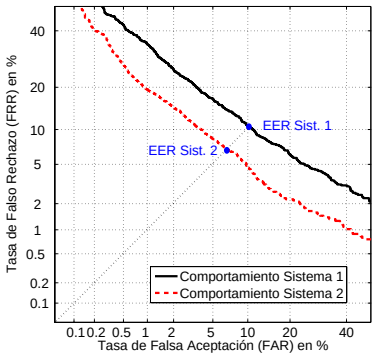

Figura 6. Ejemplo curva DET y EER

1.4 Conclusiones

El presente proyecto se ha centrado en la investigación y mejora de los sistemas de reconocimiento de locutor e idioma basados en máquinas de vectores soporte. Se ha conseguido la implementación de sistemas competitivos al nivel del estado del arte, realizando aportaciones originales y presentando unos amplios resultados.

El entrenamiento de los modelos basado en regresión, ha sido una de las aportaciones originales más importantes llevadas a cabo en el campo de reconocimiento de locutor en este proyecto. La estimación de una función que se ajustara a los datos en lugar de simplemente clasificarlos en base a su distancia al hiperplano, como hacía la clasificación, obtuvo unos resultados sensiblemente mejores. Además, se mostró como mediante el ajuste del parámetro ε es posible adaptarse a la variabilidad de los datos (variabilidad intersesión, efectos del canal, etc.). La tarea realizada en las pruebas fue la *1conv-1conv* de NIST SRE 2006, obtenido mejoras en términos de EER del 34% y 29% para género masculino y femenino respectivamente.

Una de las aportaciones originales en el campo del reconocimiento de idioma se consiguió a través de la fusión del sistema trabajando con dos parametrizaciones distintas, *MFCC* y *SDC*. El sistema global obtuvo un EER del 10.5%, sobre la tarea de 30 segundos del protocolo de evaluación NIST LRE 2005, lo que supone una mejora del 33% con respecto a los subsistemas individuales.

A parte de estas investigaciones, se ha presentado un extenso trabajo experimental en el que se ha examinado desde la influencia de ciertas variables en el comportamiento de los sistemas, hasta su rendimiento tras la implementación de distintos tipos de normalizaciones de puntuaciones y compensación de la variabilidad intersesión.

Es importante resaltar que los sistemas implementados han sido financiados por proyectos públicos y privados (véase apéndice de financiación), produciéndose una transferencia de tecnología a distintos organismos. Además, el sistema de reconocimiento de idioma fue presentado a la evaluación NIST LRE 2007 y el de locutor será incluido en NIST SRE 2008, actualmente en curso.

2. Originalidad

En este proyecto se han presentado nuevos métodos además del estado del arte de las técnicas existentes para reconocimiento de locutor e idioma. Los nuevos métodos implementados han mejorado el rendimiento de los sistemas de reconocimiento de locutor basado en SVMs. Las aportaciones de este proyecto al estado del arte y al diseño de sistemas de seguridad basados en la señal de voz pueden resumirse en las siguientes:

- **Propuesta de un nuevo método, regresión, para el entrenamiento de los modelos SVM.** Se comprueban las ventajas de dicho método frente a la clasificación clásica, dando lugar a sistemas más robustos y con un comportamiento sensiblemente mejor. Trabajo publicado en [Lopez-Moreno *et al.,* 2007].

- **Se realiza un análisis sistemático de la influencia de los parámetros de configuración en el rendimiento del sistema.** Dichos experimentos fueron orientados a encontrar la mejor configuración posible del sistema, para ello se realizaron multitud de pruebas para comprobar la influencia del coste en el entrenamiento, escalado de los datos de entrada, conjunto de desarrollo, etc.

- **Comprobación de la bondad de diversas técnicas de normalización.** Se implementaron técnicas de normalizaciones de rango *(Rank Normalization)*, además de probar técnicas de normalización de puntuaciones habituales, *T-Norm, y Z-Norm.* Técnicas que dieron lugar a una mejora en el rendimiento del sistema.

- **Compensación de la variabilidad intersesión.** Estas técnicas son una de las mayores preocupaciones de la comunidad científica a día de hoy. En el trabajo realiza una compensación en el dominio de los parámetros expandidos, mostrando los efectos de la técnica para diferentes tipos de configuración posibles.

- **Se consigue un sistema final de reconocimiento de locutor en el estado del arte.** El sistema final obtenido es el resultado de la fusión del sistema SVM basado en regresión, con la mejor configuración encontrada, y un sistema GMM-SVM (sistema explicado a lo largo del proyecto). Se mostró la incorrelación de ambos sistemas que justifica la fusión.

- **Los sistemas se prueban sobre una base de datos reconocida dentro de la literatura científica, NIST SRE 2006 y NIST LRE 2005**, accesible y utilizada en numerosos estudios científicos por diferentes grupos de investigación, proporcionando resultados que pueden ser contrastados.

En el campo del reconocimiento automático del idioma se realizaron unos avances similares a los comentados anteriormente. A continuación se detallan las aportaciones principales de este proyecto al estado del arte y al diseño de sistemas de reconocimiento automático del idioma:

- **Se investiga en el campo de la fusión de sistemas con distinta parametrización,** llegando a la conclusión de que la fusión a nivel de puntuaciones reporta unos beneficios mayores que la fusión a nivel de parámetros. La mayor parte de los trabajos realizados hasta la fecha documentan la mejora existente al concatenar el vector de parámetros MFCC junto con el SDC, en este proyecto se propone la fusión de las puntuaciones los sistemas individuales MFCC y SDC, obteniendo un comportamiento mejor. Trabajo publicado en [Toledano *et al.,* 2007].

- **Se estudia el comportamiento de la regresión en el entrenamiento de los modelos SVM de idiomas.** Gran parte de los esfuerzos realizados en este campo fueron orientados a comprobar el efecto del entrenamiento regresivo de los modelos SVM, se localizó el problema y se comenzó con una línea de investigación que trataba de subsanarlo. En estos momentos se está preparando un artículo que resuma los trabajos llevados a cabo y la problemática encontrada.

3. Resultados

En esta sección se incluirán los resultados más relevantes obtenidos a lo largo del proyecto. Como se explicó en la sección 1.3.3 los protocolos sobre los que se evalúan los sistemas son NIST SRE 2006 y NIST LRE 2005, en [Leeuwen *et al*., 2006] puede encontrarse un estudio donde se muestra que las bases de datos son significativas estadísticamente.

El protocolo de pruebas *1conv-1conv* de NIST SRE 2006 está compuesto por 353 modelos de género masculino y 462 de género femenino, lo que conlleva el entrenamiento de 815 modelos en total. La fase de test está formada por 23179 enfrentamientos para género masculino y 30656 enfrentamientos para género femenino, lo que hace un total de 53835 enfrentamientos. Todas estas cifras no tienen en cuenta modelos ni enfrentamientos adicionales necesarios para normalización de puntuaciones.

El protocolo de pruebas de 30 segundos de NIST LRE 2005 está compuesto por 7 idiomas y 25634 enfrentamientos. Al igual que en reconocimiento de locutor estas cantidades no tienen en cuenta enfrentamientos adicionales necesarios para algunos procesos de normalización de puntuaciones.

3.1 Resultados reconocimiento de locutor

En este campo se van a mostrar cuatro de los experimentos más representativos llevados a cabo con el sistema de reconocimiento de locutor basado en SVM, en las secciones 8 y 9 del proyecto original pueden encontrarse una gran variedad de experimentos más.

- Normalización de puntuaciones
- Entrenamiento de modelos basados en regresión
- Compensación de la variabilidad intersesión (NAP)
- Fusión suma SVM y GMM-SVM

Normalización de puntuaciones

La normalización de puntuaciones [Bimbot *et al.,* 2004] se emplea para realizar escalado de distribuciones y eliminar ciertas dependencias (como por ejemplo del canal) contribuyendo de ese modo a la robustez del sistema. Resulta además importante de cara a la fusión debido a que los sistemas a fusionar pueden no ser homogéneos (uno puede medir similitudes y otro distancias). Adicionalmente, situar las puntuaciones en rangos comparables permite utilizar un umbral por sistema en lugar uno específico para cada locutor.

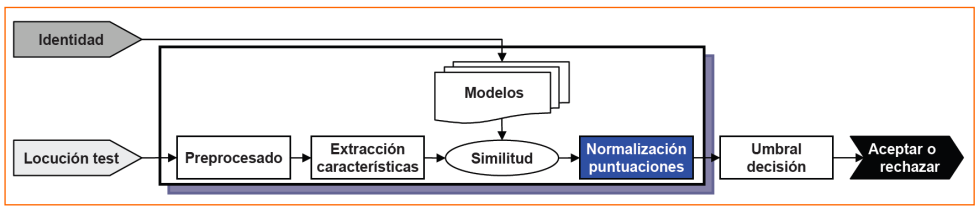

Figura 7. Esquema funcionamiento sistema SVM con normalización de puntuaciones

La Figura 8 y la Tabla 1 muestran los resultados obtenidos por el sistema sin ningún tipo normalización de puntuaciones *(raw)*, sistema base sobre el que se aplicarán todas las mejoras, y tras aplicar estas normalizaciones. En rojo y negrita se muestran los mejores resultados. Puede apreciarse como tanto el EER como el DCF mejoran con respecto al sistema base.

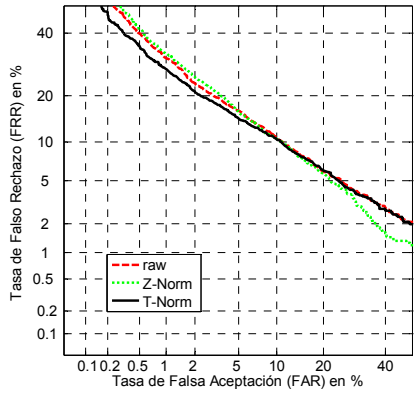

Figura 8. Curvas DET con distintas normalizaciones de puntuaciones

Norm.	Color	EER (%)	DCF
raw	Rojo --	10.42	0.041
Z-Norm	Verde …	**10.29**	0.042
T-Norm	Negro	10.42	**0.037**

Tabla 1. Comparación resultados normalización puntuaciones

Entrenamiento de modelos basados en regresión

Como se explicó anteriormente este es uno de los avances más importantes realizados en el proyecto. Los modelos entrenados con regresión *(SVR)* presentan un comportamiento más eficiente frente a posibles distorsiones de la señal de voz que los entrenados con clasificación *(SVC)*. Las Figuras 9, 10 y la Tabla 2 resumen el comportamiento de ambos sistemas, puede apreciarse una mejora del 34% y 29% en EER y del 22% y 25% en DCF para género masculino y femenino respectivamente.

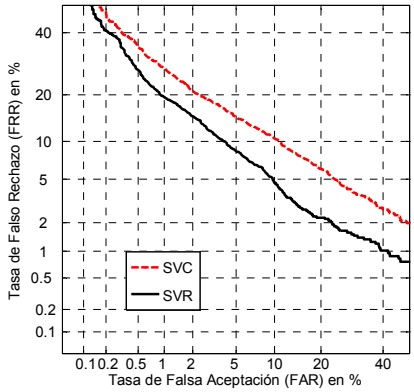

Figura 9. Curvas DET entrenamiento con regresión, género masculino

Género	Tipo SVM	Color	EER (%)	DCF
Masculino	SVC	Rojo ---	10.42	0.037
	SVR	Negro	**6.89**	**0.029**
Femenino	SVC	Rojo ---	11.98	0.048
	SVR	Negro	**8.47**	**0.035**

Tabla 2. Comparación resultados entrenamiento con regresión y clasificación

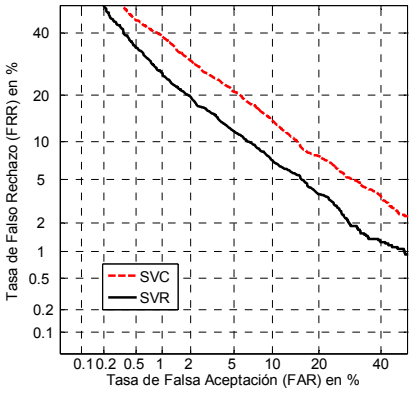

Figura 10. Curvas DET entrenamiento con regresión, género femenino

En el proyecto (apartado 8.10) se incluye un estudio detallado del valor de la variable ε en el entrenamiento. A través de esta variable es posible adaptar el entrenamiento a la variabilidad de los datos (variabilidad intersesión, efectos del canal, etc.). Los resultados incluidos en el presente documento fueron obtenidos con un valor de ε de 0.1.

Compensación de la variabilidad intersesión (NAP)

La compensación de la variabilidad intersesión es un punto clave a día de hoy en los sistemas comerciales, dada la alta variación que la señal de voz sufre debido a múltiples factores (canal, entorno, etc.). Esquemas de compensación [Solomonoff *et al.,* 2004; Solomonoff *et al.,* 2005] intentan compensar esta variabilidad y resultan muy interesantes en el campo de los SVMs.

La técnica empleada en el trabajo para compensar esta variabilidad es la ya mencionada anteriormente, NAP. Esta técnica trata de eliminar las direcciones de máxima variabilidad de la señal de voz, de esta forma evitaremos efectos no deseados

preservando la información propia del locutor. Durante la realización del proyecto se realizaron diversas pruebas con distintas configuraciones, en este resumen sólo se incluirá la mejor configuración encontrada (véase Figura 11), donde se puede apreciar una mejora del 12% en EER.

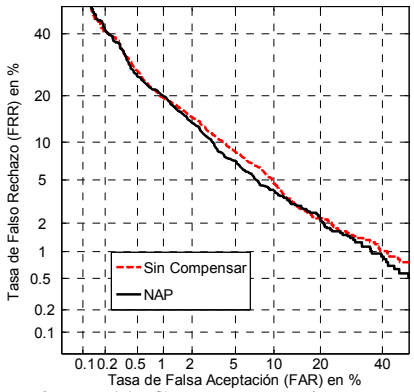

Figura 11. Curvas DET sistema con compensación NAP

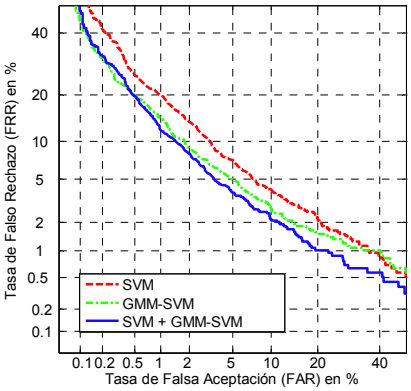

Figura 12. Curvas DET fusión sistemas SVM y GMM-SVM

Fusión suma SVM y GMM-SVM

La fusión de sistemas permite generar una única decisión final a partir de varios sistemas individuales, con un resultado final mejor que el de cada uno de los subsistemas por separado [Reynolds *et al.,* 2003b]. En el presente trabajo se incluye la fusión suma del sistema SVM basado en regresión y el sistema GMM–SVM. La Figura 12 muestra el comportamiento de los subsistemas y del sistema final, pudiendo apreciarse una mejora sustancial del EER.

3.2 Resultados reconocimiento de idioma

En este campo consta de especial importancia los resultados obtenidos con la fusión a nivel de puntuaciones del sistema basado en parametrización MFCC y del sistema basado en parametrización SDC.

De la Figura 13 y la Tabla 3 se pueden extraer dos conclusiones claras; por un lado la parametrización SDC, la más empleada en reconocimiento de idioma, presenta un comportamiento ligeramente superior a la MFCC. Por otra parte la fusión suma de ambas parametrizaciones supone una mejora del 33% en EER y 27% en DCF frente a los subsistemas individuales. Por último, resaltar que si la fusión se realiza a nivel de parámetros la mejora supone tan sólo el 5% en EER y 6% en DCF.

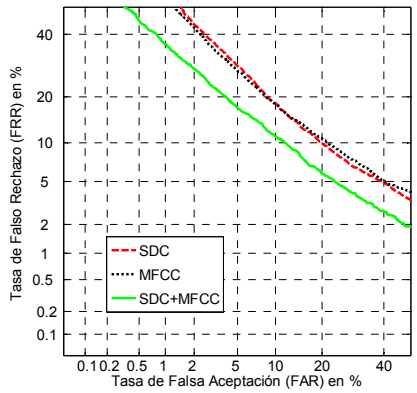

Figura 13. Curvas DET fusión sistemas con distinta parametrización

Sistema	Color	EER (%)	DCF
SDC	Rojo ---	13.84	0.061
MFCC	Negro …	14.07	0.061
SDC+MFCC	Verde	**10.55**	**0.045**

Tabla 3. Comparación resultados sistemas con distinta parametrización

4. Aplicabilidad

Hoy en día los sistemas de seguridad han cobrado una gran relevancia, en gran medida esta relevancia viene fomentada por la demanda cada vez mayor de la sociedad por este tipo de aplicaciones. Los sistemas de reconocimiento biométrico se presentan como una buena alternativa frente a los métodos clásicos de seguridad, llaves, códigos, etc. La señal de voz es un rasgo biométrico que presenta innumerables ventajas, fácil adquisición, transporte sencillo a través de redes de comunicaciones habituales: redes IP, red telefonía, etc. Estas ventajas hacen de la señal de voz un rasgo biométrico ideal para la implementación de sistemas de seguridad.

En la primera parte de este proyecto se ha investigado, propuesto e implementado un nuevo método para el reconocimiento biométrico basado en la señal de voz. En la sección 3.1 se presentó el rendimiento del nuevo método propuesto sobre un protocolo de pruebas ampliamente aceptado por la comunidad científica. Como se vio en dicha sección el rendimiento del sistema mejoraba en todos los posibles puntos de trabajo del sistema, siendo alrededor de un 30% mejor en el EER. El coste computacional del método presentado es similar al que presentan métodos habituales de reconocimiento de locutor basados en SVM. Por tanto, el sistema podría ser aplicado directamente en entornos reales fuera del laboratorio, mejorándose de esta manera el estado del arte actual en sistemas de reconocimiento de locutor basados en SVM.

El sistema desarrollado en este proyecto será incluido junto con otros sistemas elaborados por el ATVS *(Área de Tratamiento de Voz y Señal)* [ATVS], grupo de investigación donde se realizó el presente proyecto, en la evaluación NIST SRE 2008. Esta evaluación se llevará a cabo en abril del presente año. Además, cuenta con la financiación de distintos proyectos (véase apéndice de financiación), públicos y privados, a través de los cuales se producirá una transferencia de tecnología.

Además del nuevo método mencionado se han realizado multitud pruebas en las que se muestran diversas configuraciones posibles del sistema, técnicas de normalización de puntuaciones, compensación de canal, etc. Todos los resultados presentados son perfectamente reproducibles por cualquier grupo de investigación, por lo que pueden ser tenidos en cuenta a la hora de elegir la configuración y técnicas a aplicar en cualquier sistema real.

En la segunda parte del proyecto se sigue una forma de trabajo similar a la mencionada anteriormente pero en este caso para el reconocimiento automático del idioma. Este tipo de sistemas constan de una gran importancia en aplicaciones de seguridad y defensa a la hora de detectar el idioma de una conversación, también son importantes para centralitas telefónicas, sistemas de información, banca telefónica, etc. donde la detección del idioma del locutor puede ser crucial a la hora de tratar convenientemente una llamada. De esta forma las llamadas pueden ser redirigidas a un operador adecuado o cambiar el idioma en el que sistema automático proporcione la información.

Una de las propuestas más importantes en este campo ha sido la fusión a nivel de puntuaciones de dos sistemas implementados, sistema de reconocimiento de idioma basado en una parametrización MFCC y sistema de reconocimiento de idioma basado en una parametrización SDC. En un sistema real, donde los modelos de los idiomas ya

están entrenados, el proceso sería muy simple, tendrían que realizarse los enfrentamientos con los dos tipos de modelos para posteriormente sumar las puntuaciones obtenidas. Como se mostró en la sección 3.2 el rendimiento del sistema mejoraría un 33%, reduciendo de este modo los errores cometidos a costa de incrementar ligeramente el coste computacional.

El sistema desarrollado, junto con otros desarrollados por el grupo ATVS, fueron presentados a la evaluación NIST LRE 2007 [ATVS-UAM NIST LRE 2007]. Donde fue comparado con sistemas de reconocimiento de idioma de distintos grupos de investigación a nivel mundial, obteniendo unos resultados a nivel del estado del arte. Al igual que sucedía en reconocimiento de locutor, los trabajos realizados cuentan con la financiación de distintos proyectos, públicos y privados, a través de los cuales se producirá una transferencia de tecnología.

Por último, al igual que se comentó en el caso del reconocimiento de locutor, la gran cantidad de experimentos llevados a cabo sobre un protocolo de pruebas ampliamente aceptado por al comunidad científica hace que los resultados presentados sean reproducibles por cualquier grupo de investigación, ayudando de este modo a elegir distintas configuraciones del sistema y técnicas de normalización, compensación, etc.

Apéndice I: Referencias

ATVS, Área de Tratamiento de Voz y Señal, Escuela Politécnica Superior, Universidad Autónoma de Madrid. http://atvs.ii.uam.es

F. Bimbot et al. A tutorial on text-independent speaker verification. *Journal on Applied Signal Processing*, N. 4, pp. 430-451, 2004.

W. M. Campbell. Generalized linear discriminate sequence kernels for speaker recognition. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 1, pp. 161-164, 2002.

W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, y P. A. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, Vol. 20, N. 2-3, pp. 210-229, 2006b.

J. R. Deller, J. H. L. Hansen y J. G. Proakis. Discrete-time processing of speech signals. *Wiley-IEEE Press*, Septiembre 1999.

S. Furui. Cepstral Analysis technique for automatic speaker verification. *IEEE Transactions on acoustics, speech and signal processing,* Vol. ASSP-29, N. 2, Abril 1981.

H. Hermansky y N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing 2*, pp. 578-589, Octubre 1994.

LDC, *(Linguistic Data Consortium)*. Descripción de distintas bases de datos. www.ldc.upenn.edu

A. Martin, G. Doddington, T. Kamm, M. Ordowski y M. Przybocki. The DET curve in assessment of decision task performance. *Proc. EuroSpeech*, pp. 1895-1898, 1997.

D. van Leeuwen, A. Martin, M. Przybocki y J. Bouten. The NIST 2004 and TNO/NFI Speaker Recognition Evaluations. *Computer Speech and Language*, Vol. 20, pp. 128-158, 2006.

I. Lopez-Moreno, I. Mateos-García, D. Ramos y J. Gonzalez-Rodriguez. Support vector regression for speaker verification. In *Proc. Interspeech*, pp. 306-309, Antwerp, Belgium, August 2007.

NIST. Grupo de trabajo de NIST dedicado a temas relacionados con la señal de voz, http://www.nist.gov/speech

NIST LRE. Descripciones de las distintas evaluaciones NIST de idioma, http://www.nist.gov/speech/tests/lang

NIST SRE. Descripciones de las distintas evaluaciones NIST de locutor, http://www.nist.gov/speech/tests/spk

J. Pelecanos y S. Sridharan. Feature warping for robust speaker verification. *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pp. 213-218, 2001.

D. A. Reynolds et al. Supersid project: Exploiting high-level information for high accuracy speaker recognition. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 4, pp 784-787, Abril 2003.

A. Solomonoff, C. Quillen, y W.M. Campbell. Channel compensations for SVM speaker recognition. *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey),* pp. 57-62, 2004.

A. Solomonoff, W.M. Campbell y I. Boardman. Advances channel compensations for SVM speaker recognition. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Vol. 1, pp. 629-632, 2005.

D. T. Toledano, I. Lopez-Moreno, I. Mateos-García, A. Abejon-Gonzalez, D. Ramos y J. Gonzalez-Rodriguez. ATVS-UAM NIST LRE 2007 System Description.

P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds y J. R. Deller Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstrum. *Proc. International Conference on Spoken Language (ICSLP)*, 2002.

Apéndice II: Financiación

El proyecto ha contado con la siguiente financiación:

Proyecto: Reconocimiento eficiente multinivel del locutor y del idioma para su uso policial operativo, como evidencia forense y como control de acceso.
Entidad financiadora: Ministerio de Educación y Ciencia, TEC2006-13170-C02-01/TCM.

Proyecto: Fusión de Enfoques Acústicos Generativos y Discriminativos con Niveles Superiores en Identificación de Idioma y Locutor.
Entidad financiadora: Ministerio de Defensa.

Proyecto: Calibración de relaciones de verosimilitud y combinación de evidencias en acústica forense.
Entidad financiadora: Dirección General de la Guardia Civil, Ministerio del Interior.

Proyecto: Convenio para la transferencia de tecnología de reconocimiento de locutor.
Entidad financiadora: Agnitio.

Apéndice III: Publicaciones

El proyecto ha dado lugar a la publicación del trabajo *"Support vector regresión for speaker verification"* en el congreso internacional de las tecnologías del habla, Interspeech 2007, anexado a continuación y publicado bajo la siguiente referencia:

> I. Lopez-Moreno, I. Mateos-García, D. Ramos y J. Gonzalez-Rodriguez. "Support vector regression for speaker verification". In *Proc. Interspeech*, pp. 306-309, Antwerp, Belgium, August 2007.

Adicionalmente, ha contribuido a la publicación del trabajo *"Improved language recognition using better phonetic decoders and fusion with MFCC and SDC features"* en el mismo congreso indicado anteriormente. El artículo se anexa a continuación y está recogido bajo la referencia:

> D. T. Toledano, J. Gonzalez-Dominguez, A. Abejón-Gonzalez, D. Spada, I. Mateos-García y J. Gonzalez-Rodriguez. "Improved language recognition using better phonetic decoders and fusion with MFCC and SDC features". In *Proc. Interspeech*, pp. 194-197, Antwerp, Belgium, August 2007.

Por último, el sistema de reconocimiento de idioma tomó parte, junto con otros sistemas del grupo en la evaluación más reciente de NIST en este campo, NIST LRE 2007. La descripción del sistema, *"ATVS-UAM NIST LRE 2007 System Description"* se incluye a continuación y responde a la siguiente referencia:

> D. T. Toledano, I. Lopez-Moreno, I. Mateos-García, A. Abejon-Gonzalez, D. Ramos y J. Gonzalez-Rodriguez. ATVS-UAM NIST LRE 2007 System Description.