RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
Reconocimiento de Patrones

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor

José Ramón Calvo, Rafael Fernández, Gabriel Hernández

RT_008 **Febrero 2008**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor

José Ramón Calvo, Rafael Fernández, Gabriel Hernández

RT_008 **Febrero 2008**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor

José Ramón Calvo, Rafael Fernández, Gabriel Hernández

Centro de Aplicaciones de Tecnología de Avanzada, 7a #21812 e/ 218 y 222, Siboney, Playa, Habana, Cuba
jcalvo@cenatav.co.cu

RT_008 CENATAV
Fecha del camera-ready: Febrero 2008

Resumen: El reconocimiento automático del locutor es una de las áreas de la biometría que más importancia ha adquirido en los últimos años. En este trabajo se presentan y analizan las principales técnicas de extracción y selección de rasgos acústicos así como las de clasificación para el reconocimiento del locutor. Se muestran además resultados de experimentos realizados sobre una base de datos española y se establecen los principales problemas a resolver en estos temas así como las líneas de trabajo futuro.

Palabras clave: reconocimiento del locutor, extracción de rasgos, selección de rasgos, clasificación.

Abstract: The growing impact of the applications based on automatic speaker recognition has turned this biometric area into one of the most important in the last years. The most relevant feature extraction and selection techniques along with the classification methods for speaker recognition are presented. In this sense, results of experiments over a Spanish database are shown. The main problems in the state of the art and the future works are summarized.

Keywords: speaker recognition, feature extraction, feature selection, classification.

Índice general

Índice de figuras

Índice de cuadros

1. Introducción

El reconocimiento automático de una persona por su voz, o reconocimiento automático del locutor, es actualmente un área de investigación y de desarrollo de aplicaciones de gran importancia.

Las primeras aproximaciones documentadas del reconocimiento automático de locutores se producen a principios de los años setenta. Los ingenieros de la Bell [1], Bishnu Atal [2] y Aaron Rosenberg y Sanbur (1975) [3] publican sus primeros estudios utilizando ya, como base de extracción de datos, coeficientes cesptrum y coeficientes de predicción linear o LPC (por sus siglas en inglés). En esta misma época, son evaluados diversos parámetros acústicos del habla para su utilización en el diseño de sistemas de reconocimiento automático: en 1972, Wolf [4] analiza combinaciones de hasta veintisiete medidas extraídas de consonantes nasales, espectros de vocales y la frecuencia fundamental. En 1973 Su y Fu [5] utilizan como informacioneseficientes los espectros de consonantes nasales. Li y Hughes en 1974 [6], toman como referencia matrices de correlación referidas a fragmentos de habla continua. La mayoría de estos primeros intentos estaban estrechamente ligados a comparaciones dependientes de texto, aunque igualmente se reportan estudios forenses de reconocimiento automático independientes de texto [7].

Esta rama tan antigua en el establecimiento de sus principios teóricos como el reconocimiento automático del habla, ha tenido, sin embargo, poca atención y como consecuencia, un desarrollo menor. Ha sido en estos últimos años cuando, a la luz de los nuevos e importantes campos de aplicación surgidos –como la biometría– que su desarrollo ha sido mayor.

1.1. Métodos básicos de reconocimiento automático del locutor

Dentro del reconocimiento automático del locutor (RAL) podemos distinguir dos tareas diferenciadas:

- **Verificación automática del locutor (VAL):** El objetivo es verificar la identidad reclamada por el locutor, o sea, tenemos un individuo que dice ser alguien y una muestra de su voz, la tarea a realizar es ver si ambas coinciden o no; la respuesta del sistema será, por lo tanto, binaria: identidad aceptada o rechazada. Hay diversas formas de medir la efectividad de este tipo de sistemas, una de las más utilizadas es la denominada tasa de EER[1]: es el error del sistema cuando el umbral decisión es tal que el porcentaje de falsas aceptaciones es igual al de falsos rechazos (Fig. 1) entonces, si la salida del sistema es menor que el umbral de decisión[2] la identidad reclamada por el cliente es rechazada, en caso contrario, será aceptada.
- **Identificación automática del locutor (IAL):** Aquí el objetivo es, dada una muestra de voz, señalar, dentro de un grupo de personas, la identidad de su "propietario". Hablamos de IAL de Grupo Cerrado si el locutor desconocido es con certeza uno de los del grupo, y de IAL de Grupo Abierto si existe la posibilidad de que el locutor

[1] Equal error rate (EER)

[2] El umbral en este caso es una medida de la similitud entre un conjunto de rasgos de prueba y un modelo perteneciente a la base de datos.

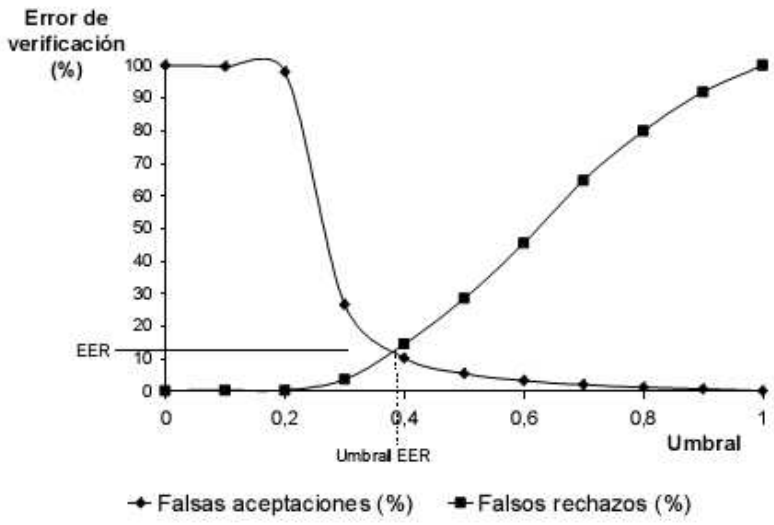


Figura 1. Curvas de error. Falsas aceptaciones (FA) y falsos rechazos (FR). FA: la entrada pertenece a un impostor y es falsamente aceptada. FR: la entrada pertenece al cliente y es falsamente rechazada

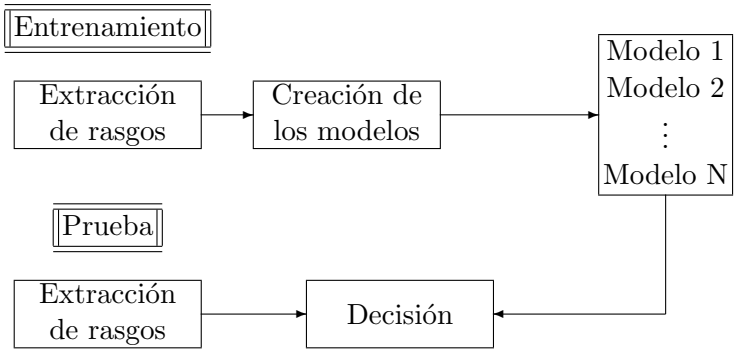
pueda ser alguien ajeno a ese grupo de personas. En el primer caso la respuesta del sistema será siempre una identidad, mientras que en el segundo existe la posibilidad de la respuesta sea locutor rechazado al no pertenecer al grupo de personas a identificar.

Sadaoki Furui –ingeniero de NTT *Human Interface Laboratories*, Tokio– define estos conceptos de la siguiente manera:

- *Reconocimiento de locutores*: Todo proceso automático de reconocimiento de hablantes basado en la información individual incluida en la señal de habla. Dicho proceso se divide en Identificación y Verificación de hablantes.
- *Identificación de hablantes*: Proceso por el que se determina a quién pertenece la muestra anónima aportada, de entre un número de muestras registradas pertenecientes a distintos hablantes.
- *Verificación de hablantes*: Proceso de aceptación o rechazo de identidad a través de voz, solicitado por un hablante.

En relación con dichos conceptos, Furui en 1994 [8] puntualiza: "*...la diferencia fundamental entre identificación y verificación es el número de decisiones alternativas. En identificación, el número de decisiones alternativas es igual al número de sujetos de la población que conforma la base de datos, mientras que en verificación sólo existen dos decisiones alternativas, aceptar o rechazar, con independencia de la talla de la población.*"

Un sistema automático para la identificación del locutor presenta el siguiente esquema general [9]:



En la etapa de entrenamiento [9] se obtienen los modelos y se establecen los umbrales correspondientes a cada locutor perteneciente a la base de datos previamente elaborada.

En la etapa de prueba se comparan los modelos de la base con las muestras de voz del sospechoso, lo que permitirá tomar decisiones acerca de la identidad del mismo.

En ambas etapas es necesario una adquisición adecuada de las muestras de voces, mejorar su calidad –sin afectar su inteligibilidad ni sus rasgos para el procesamiento posterior– y extraer de forma automática los rasgos acústicos.

La identidad de un locutor [9] puede llevarse a cabo desde información de bajo nivel, como los rasgos acústicos y fonéticos del habla del locutor, hasta información de alto nivel, como las peculiaridades psico-lingüísticas del mismo. Los humanos reconocen a los locutores fundamentalmente por el procesamiento en el cerebro de la información de alto nivel, mientras que los sistemas automáticos reconocen a los locutores a partir de la información de bajo nivel extraída en rasgos acústicos temporales y espectrales, de corto y largo término. Los rasgos acústicos de la voz para la identificación automática del locutor deben presentar las características ideales siguientes [10]:

- Que tengan alto poder discriminativo y sean fácilmente medibles,
- que ocurran naturalmente y frecuentemente,
- que varíen lo más posible entre locutores pero que sean consistentes en cada locutor,
- que no cambien en el tiempo o se afecten por la salud del locutor,
- que no sean modificables por un esfuerzo conciente del locutor de disfrazar su voz,
- que sean capaces de crear un modelo estadístico de manera fácil y con propiedades invariantes en un amplio rango de ambientes de locución.

Este reporte está dirigido a recoger el estado del arte de los métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor. Está dividido en tres capítulos, que tratan de los métodos descritos.

2. Principales métodos de extracción de rasgos acústicos del habla

El proceso de extracción de rasgos consiste en una transformación de un espacio de alta dimensionalidad a otro de menor dimensionalidad. Es decir, es una correspondencia $f : \mathbb{R}^N \rightarrow \mathbb{R}^d$ donde $d \ll N$. Esta es una operación necesaria por dos motivos fundamentales:

en primer lugar, para reducir el volumen de cálculo, y en segundo lugar, para evitar el problema conocido como la *maldición* de la dimensionalidad [10], que básicamente consiste en que el número de vectores de entrenamiento necesarios para una caracterización robusta del locutor crece exponencialmente con la dimensión del espacio. A continuación se analizan los principales métodos de extracción de rasgos.

2.1. Bancos de filtros

La extracción de rasgos por procesado en sub-bandas, conocidos como *bancos de filtros* (*filterbanks* en inglés), es un término genérico que se refiere a los métodos que procesan una señal en múltiples bandas de frecuencia y fueron propuestos por [11]. Se diseña un banco de filtro en el dominio de tiempo por medio de ecuaciones recursivas, o en el dominio de la frecuencia multiplicando el espectro de la señal con la respuesta de magnitud de cada sub-banda del filtro. En el diseño en el dominio del tiempo [12], la extracción de rasgos se hace para la señal en cada sub-banda usando el proceso de análisis trama a trama con las mismas técnicas que si fuese la señal en toda la banda, trayendo como consecuencia práctica que la resolución de cada sub-banda puede controlarse más fácilmente que en el proceso de toda la banda. En el diseño en el dominio de la frecuencia, son ejemplo los bancos de filtros distorsionados en frecuencias *Mel* y *Bark*, para simular la percepción del oido humano [13].

Este método tiene la ventaja sobre otras representaciones espectrales, que los rasgos tienen una interpretación física directa, por ejemplo, conociendo *a priori* el poder discriminativo de cada sub-banda, se puede establecer un peso apropiado a cada una [14,15], o si alguna de las sub-bandas esta contaminada por ruido, pueden usarse las no contaminadas [16].

Según [10,17] existen dos formas de utilizar los bancos de filtros en el reconocimiento del locutor: realizando la fusión a nivel de rasgo (fusión de entrada), o la fusión a nivel del clasificador (fusión de salida). En la primera se combinan los rasgos obtenidos en cada sub-banda en un solo vector de rasgos M-dimensional y se entrena un solo modelo de locutor. En la segunda, los rasgos de cada sub-banda son considerados independientemente y para cada sub-banda, se crea un modelo separado.

Pueden considerarse ventajas de los bancos de filtros en el reconocimiento del locutor:

- el sistema puede ser robusto en el caso de habla afectada por ruido en un limitado número de sub-bandas,
- podrían aplicarse diferentes técnicas del reconocimiento a diferentes sub-bandas,
- la aproximación de las sub-bandas puede aprovecharse de las arquitecturas paralelas,
- pueden localizarse cuáles son las sub-bandas más específicas a cada locutor. En general, las sub-bandas de baja frecuencia (por debajo de 600 Hz) y las sub-bandas de alta frecuencia (por encima de 3000 Hz) son más específicas al locutor que las sub-bandas de media-frecuencia [16].

Por último, la eficacia del método los bancos de filtros depende significativamente de varios factores:

- la arquitectura del sistema: selección de las sub-bandas más críticas para el reconocimiento (incremento de peso en la decisión); la división óptima de la banda total de frecuencias (número y tamaño de sub-bandas),
- la recombinación de la salida de cada reconocedor por sub-banda: nivel de recombinación, estrategias de recombinación, fusión de decisiones múltiples,
- los rasgos a usar para el reconocimiento del locutor deben ser seleccionados. Los rasgos que se usan para el reconocimiento del habla, por lo general no son recomendables para el reconocimiento del locutor. De igual modo, algunos rasgos apropiados para el reconocimiento usando todo el espectro, pueden ser ineficientes al aplicar un procesamiento multiespectral [18].

En [12] se estudia el reconocimiento del locutor en presencia de ruido gaussiano de banda estrecha mediante el procesamiento por subbandas, usando un banco de filtros pasabandas *Butterworth* de sexto orden, en escala *Mel*. Aplican cadenas ocultas de Markov (HMM, por sus siglas en inglés) y obtienen la probabilidad $p(\boldsymbol{x}\,|\,\omega(n,s))$ de que los datos de prueba $\boldsymbol{x}$ hallan sido producidos por el locutor s filtrado por la subbanda n, representado con el modelo $\omega(n,s)$. El locutor identificado i se obtiene aplicando la siguiente regla de fusión de las decisiones (*decision fusion rule*):

$$i = \arg\max_{s} \sum_{n=1}^{N} \log p(\boldsymbol{x}\,|\,\omega(n,s)) \quad 1 \leq s \leq S. \tag{1}$$

La identificación tuvo sus peores resultados para ruido con frecuencia promedio de $1000 - 1500$ Hz. A pesar de que no hicieron experimentos con ruido no estacionario, consideran que en presencia de éste, los resultados no empeoren.

2.2. Coeficientes de predicción lineal

El tracto vocal puede considerarse como un tubo acústico conformado por cilindros de diferentes secciones transversales, sin pérdidas ni ramificaciones, con una onda plana de sonido propagándose y reflejándose en toda su extensión desde la glotis hasta los labios, su efecto en la excitación glotal es provocarle una serie de resonancias; es decir, el tracto vocal puede modelarse como un filtro todo-polo, siendo una buena aproximación para muchos sonidos del habla en condiciones acústicas favorables:

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k\, z^{-k}}, \tag{2}$$

donde $S(z)$ y $E(z)$ son las transformadas–z de la señal de voz $s[n]$ y de la excitación producido en la glotis $Ge[n]$, con G un coeficiente de ganancia.

Aplicando en 2 la transformada–z inversa se obtiene:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Ge[n]. \tag{3}$$

El método conocido como de coeficientes de predicción lineal (LPC) o de modelación autorregresiva, propuesto en [19] para la codificación de la voz y generalizado en [20] para

el análisis de la voz, ajusta los parámetros del filtro todo-polo al espectro del habla. Con un número suficiente de coeficientes, el método LPC puede considerarse una aproximación adecuada a la estructura espectral de muchos tipos de sonidos. Dicho método fue utilizado en [2] para la obtención de rasgos acústicos en el reconocimiento del locutor, permitiendo separar en los sonidos sonoros, la información dependiente del locutor (forma, longitud del tracto y comportamiento de los formantes) de la excitación, brindando una información combinada sobre la frecuencia y ancho de banda de los formantes y el comportamiento de la onda glotal.

La idea del método LPC es que muestras adyacentes en la señal de la voz están altamente correlacionadas, y por tanto, el comportamiento de esta en un momento dado, puede predecirse de cierta cantidad de muestras anteriores, haciendo la siguiente aproximación:

$$\tilde{s}[n] \approx \sum_{k=1}^{p} a[k]s[n-k], \tag{4}$$

donde p es el *orden* de la predicción. Luego, el objetivo del método LPC consiste en determinar los *coeficientes de predicción* $\{a[k] \,|\, k = 1, \ldots, p\}$ tal que el error de predicción promedio o residuo sea mínimo. Este error, para la n-ésima muestra, viene dado por la diferencia entre la muestra actual y el valor predicho por el modelo:

$$err[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} a[k]s[n-k]. \tag{5}$$

El problema de optimización se reduce a solucionar las llamadas ecuaciones autorregresivas de Yule-Walker [21,22]. En dependencia del intervalo de minimización del error hay dos métodos para resolver las ecuaciones autorregresivas: covarianza o autocorrelación. De acuerdo con [21], para habla no sonora, ambos métodos dan resultados similares pero para habla sonora, el método de la covarianza es más exacto. Sin embargo, según [23], el método de autocorrelación es el preferido, dado que es computacionalmente más eficiente y garantiza siempre un filtro estable.

Al obtenerse los coeficientes del filtro todo-polo, la función transferencial del mismo representa el comportamiento espectral del tracto vocal que modula la excitación glotal, reflejando la distribución espectral suavizada del sonido en el intervalo de muestras tomado para llevar a cabo la predicción [24].

En la práctica, los LPC son altamente correlacionados [10] y raramente se utilizan como rasgos por sí mismos, de los posibles coeficientes que se pueden obtener a partir del modelado de la voz como un proceso autorregresivo, los coeficientes cepstrales, conocidos como LPC-cepstrales o LPCC, han sido siempre utilizados por muchos sistemas de reconocimiento del locutor [25,26,27], al estar menos correlacionados. Una envolvente espectral reconstruida a partir de coeficientes cepstrales es más suavizada que una reconstruida a partir de LPC, brindando una representación más estable entre una repetición y otra de las expresiones de un locutor dado [28]. El método LPC proporciona una alternativa sencilla al método de derivar los coeficientes cepstrales de los LPC a partir de expresiones recursivas

propuestas por [25] y perfeccionadas por [21]:

$$\hat{c}[n] = \begin{cases} 0 & n < 0 \\ \ln G & n = 0 \\ a[n] + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) \hat{c}[k] a[n-k] & 0 < n \leq p \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) \hat{c}[k] a[n-k] & n > p \, . \end{cases} \tag{6}$$

No obstante, debe señalarse que los LPCC se derivan de los LPC, asumiendo el mismo modelo de filtro todo-polo, y que los LPCC no son iguales a los coeficientes cepstrales derivados directamente del espectro. Se han aplicado otras transformaciones sobre los LPC que dan lugar a distintas familias de coeficientes menos correlacionados, susceptibles de ser utilizadas en tareas de reconocimiento del locutor y del habla, con mayor o menor acierto [29]:

- *Coeficientes de reflexión* (RC, por sus siglas en inglés): Si se utiliza el método de auto-correlación de Durbin-Levinson, para resolver las ecuaciones auto-regresivas, los coeficientes de reflexión son las variables intermedias en la recursión, aunque también pueden obtenerse a partir de los LPC usando una recursión hacia atrás [24].
- *Coeficientes de razón logarítmica de áreas* (LAR, por sus siglas en inglés): el modelo asumido es el de un tubo acústico sin pérdidas, conformado por cilindros de diferentes secciones transversales en cuyas uniones se producen desacoples de impedancia que provocan las reflexiones. Los RC representan el por ciento de reflexión en dichas discontinuidades. Si el habla se pre-enfatiza antes de hacer el análisis LPC, las áreas de las secciones transversales obtenidas, son similares a las de la conformación del tracto vocal para la producción del habla que se analiza [24]. Los coeficientes LAR son el logaritmo de la razón entre áreas adyacentes del tracto vocal modelado.
- *Coeficientes de reflexión ArcSin*: para evitar posibles singularidades en los coeficientes LAR y asegurar una sensitividad espectral uniforme, se calculan los *ArcSin* de los RC.
- *Coeficientes de pares de líneas espectrales* (LSP, por sus siglas en inglés): fueron introducidos por [30] y se obtienen a partir del mapeo de los polos del filtro todo-polo sobre el círculo de radio unitario, a través de dos polinomios auxiliares [31,21], donde los pares de líneas espectrales LSP, se corresponden con las fases (ángulos) de cada uno de los ceros de los polinomios auxiliares.

Los LSP han sido utilizados en reconocimiento del locutor [29,32,33], siendo la familia de coeficientes LPC que mejores resultados ha aportado. En particular, en [32] se experimentó con la media y la diferencia de coeficientes LSP adyacentes, que correlacionan con la frecuencia y el ancho de banda de los formantes, con mejores resultados que los obtenidos con los LPCC. No obstante, según [2] y [24], el modelo del cual se obtienen los coeficientes LPC presenta las siguientes limitantes, que constituyen inconvenientes en su aplicación:

- Los pulsos glotales no tienen una estructura espectral plana,
- el tracto vocal no está compuesto solo por cilindros,

- la cavidad nasal constituye un pasaje adicional,
- algunos sonidos se generan cerca de los labios como los fricativos sordos.

Además, los LPC presentan problemas ante la degradación del habla producto del ruido y ante cambios en el canal [28], que los hacen poco útiles en ambientes reales. La *Predicción Lineal Perceptual* (PLP) propuesta en [34], combina los LPC con el método de bancos de filtros explotando principios de percepción sico-acústica, incluyendo el análisis en bandas *Bark*, el pre-énfasis de igual sonoridad y la relación de intensidad de sonoridad. La PLP y sus variantes han sido usados en aplicaciones de reconocimiento del locutor [35,36,37]. Según [37], los coeficientes PLP sobrepasan a los LPC para todas las condiciones de canal y ruido evaluadas, e incluso superan a los coeficientes cepstrales derivados directamente del espectro.

2.3. Rasgos cepstrales obtenidos del espectro

La representación de la voz utilizando su espectro de potencia a corto término asume la cuasi - estacionariedad de la voz en segmentos de tiempo entre 20 y 30 ms, lo que permite obtener su comportamiento espectral en el segmento aplicando la transformada de Fourier. La obtención continua y solapada de los espectros a corto término de la voz da lugar al espectrograma, esta técnica fue utilizada por primera vez en [38] y se ha generalizado dicha representación como herramienta de análisis de la voz. Los rasgos cepstrales pueden obtenerse también a partir del espectro de potencia a corto término de la señal del habla (*Short-Time Fourier Transform*, STFT). Estos fueron propuestos en [39] y constituyen el conjunto de rasgos de uso más común actualmente en reconocimiento del habla y del locutor. Dichos rasgos, representados en escala logarítmica y distorsionados en frecuencia con escala *Mel*, son conocidos por *Coeficientes Cepstrales en escala Mel* (MFCC) [10,21,28,29].

La representación logarítmica del espectro de potencia tiene las siguientes ventajas:

- Cuando la ganancia de la señal varía, la forma del espectro se preserva y solo se desplaza en amplitud.
- Un filtrado lineal causado por la acústica del local o por variaciones en la línea telefónica, tiene efectos convolucionales en la forma de onda y multiplicativos en el espectro de potencia, reflejándose como adiciones en el logaritmo del espectro de potencia.
- La señal sonora puede modelarse como la convolución de una excitación cuasi-periódica con un filtro variante en el tiempo que representa al tracto vocal, ambas componentes son fáciles de separar en el dominio de la potencia logarítmica, donde se suman.
- La distribución estadística del espectro en el dominio logarítmico tiene propiedades no presentes en el espectro de potencia lineal, que son convenientes en el reconocimiento del locutor y del habla.

El *cepstrum* de la voz se define como la transformada de Fourier inversa del logaritmo del espectro de potencia a corto término, y se utiliza para caracterizar tanto sonidos sonoros como sordos, con buenos resultados en la práctica tanto en el reconocimiento del habla como del locutor [21]. La señal de voz se procesa con un filtro de pre-énfasis $H(z) = 1 - \alpha z^{-1}$, que brinda un incremento de ganancia de 6 dB/octava, obteniendo un espectro promedio del habla relativamente plano. El espectro de potencia a corto término se obtiene tomando

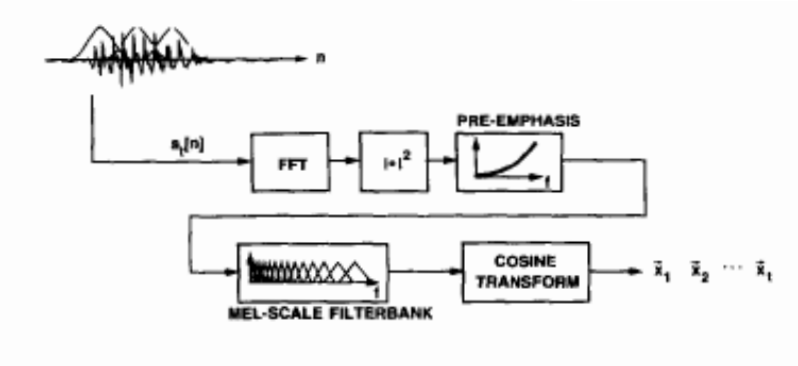


Figura 2. Esquema de extracción de los rasgos cepstrales en escala *Mel*

porciones solapadas sucesivas de la señal, típicamente entre 20 y 30 ms, a las cuales se le aplica una ventana que suaviza los extremos, y se aplica la transformada de Fourier:

$$X_a[k] = \sum_{n=0}^{N-1} x_a[n] e^{-2\pi jnk/N}, \tag{7}$$

donde $x_a[n] = x[n]w[n-a]$ con $w[n] \neq 0$ para $0 \leq n \leq N-1$.

La cóclea del oído humano realiza un análisis espectral en una escala no lineal (escala *Bark* o *Mel*), que es lineal hasta 1000 Hz y aproximadamente logarítmica después de este valor de frecuencia [39]. Por tal motivo, al extraer los rasgos espectrales a corto término, es comúnefectuar una distorsión en frecuencia después del cálculo espectral, el espectro obtenido se procesa con un banco de filtros de M bandas de frecuencia, de acuerdo a la escala *Mel*, definido –para el m-ésimo filtro– como sigue:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{k-f[m-1]}{f[m]-f[m-1]} & f[m-1] \leq k \leq f[m] \\ \frac{f[m+1]-k}{f[m+1]-f[m]} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \,, \end{cases} \tag{8}$$

los cuales satisfacen la condición $\sum_{m=0}^{M-1} H_m[k] = 1$. Aquí $f[m]$ representa la m-ésima frecuencia en la escala *Mel*:

$$f[m] = \left(\frac{N}{F}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{9}$$

donde:

$$B^{-1}(b) = 700\left(\exp\left(\frac{b}{1125}\right) - 1\right) \tag{10}$$

Posteriormente, se obtiene el logaritmo de dicho espectro de potencia distorsionado en escala *Mel*:

$$S[m] = \ln\left(\sum_{k=0}^{N-1} \left|X_a[k]\right|^2 H_m[k]\right), \qquad 0 \leq m < M. \tag{11}$$

Los MFCC vienen dados por la Transformada Discreta del Coseno (DCT) de las $S[m]$:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left[\frac{\pi}{M}\left(m + \frac{1}{2}\right) n\right], \qquad 0 \leq n < M, \tag{12}$$

donde M varía en dependencia de la implementación entre 24 y 40. Típicamente, se escogen 12 coeficientes cepstrales para el reconocimiento del locutor (de los que se excluye $c[0]$, que sólo depende de la intensidad). La aplicación de la DCT reduce la correlación entre los coeficientes, por lo que su matriz de covarianza puede considerarse diagonal [40]. Los coeficientes cepstrales más bajos representan los cambios lentos del espectro provenientes del tracto vocal y la inclinación espectral, mientras que los coeficientes mas altos representan las componentes de rápida variación del espectro, como la vibración de las cuerdas vocales en sonidos sonoros [10]. Los rasgos PLP son más robustos que los MFCC en ambientes ruidosos y condiciones incompatibles del canal. Por otra parte, los MFCC tienen el beneficio de ser bien modelados por una combinación lineal de densidades gaussianas como las usadas en el modelo de mezclas gaussianas [41] y brindan buenos resultados en sistemas de reconocimiento del locutor y del habla.

2.4. Rasgos dinámicos

Los rasgos espectrales son una "fotografía" del espectro en un cierto instante de tiempo, asumiendo que representan una señal estacionaria a corto término, sin información sobre su comportamiento en el tiempo. Al hablar, los órganos articulatorios están cambiando su forma continuamente, este movimiento se refleja en el espectro en los cambios en las frecuencias y anchos de banda de los formantes, pudiendo constituir un elemento identificativo del locutor [10]. La información dinámica de los rasgos espectrales o rasgos *delta* (Δ) [21,23] se estima calculando las derivadas de los rasgos en el tiempo y anexando dichas derivadas al vector de rasgos, elevando su dimensionalidad, lo que requiere un mayor volumen de datos para el entrenamiento de los clasificadores:

$$\Delta \boldsymbol{x}_{i,k} = \frac{\sum_{m=-M}^{M} m\, \boldsymbol{x}_{i,k+m}}{\sum_{m=-M}^{M} m^2}, \tag{13}$$

aquí M es una función –lineal y decreciente– del índice del rasgo en cuestión, ya que los rasgos cepstrales de mayor orden varían más rápidamente [42].

Comúnmente se estiman también las derivadas en el tiempo de los rasgos Δ, conocidas como rasgos *delta-delta* ($\Delta\Delta$), y se anexan al vector de rasgos, creciendo aún más la dimensionalidad del espacio de rasgos. En lugar de anexar los rasgos, en [43] se propone como alternativa un método que combina la salida de diferentes clasificadores de rasgos. Los rasgos *delta* y *delta-delta* se pueden utilizar sobre cualquier rasgo espectral a corto término,

especialmente los rasgos cepstrales y sus variantes [40,26,43]. Para estimar las derivadas puede aplicarse la derivación o el ajuste de una expresión polinomial. La diferenciación es simple pero actúa como un filtro paso-alto en el dominio del tiempo y tiende a amplificar el ruido, siendo más conveniente el ajuste de una curva polinomial a la trayectoria en tiempo del rasgo. Esto representa la derivada de la trayectoria en tiempo filtrada paso-bajo [23]. En ambos métodos se requiere un adecuado número de tramas de tiempo para obtener el mejor estimado de la dinámica cepstral; pocas tramas dan un estimado ruidoso, a medida que se aumenten las tramas, el estimado se suaviza, lo que tampoco es conveniente. Según el orden del rasgo cepstral, así será la cantidad de tramas necesarias, pues cada rasgo tiene su propia dinámica temporal [10].

2.5. Rasgos cepstrales delta desplazados (SDC)

Este método es muy utilizado –según lo reportado en la literatura– en reconocimiento del lenguaje. El uso de los rasgos SDC (por sus siglas en inglés) permite calcular un vector de rasgos pseudo-prosódico de una forma bastante sencilla sin la necesidad de encontrar un modelo de la estructura prosódica de la señal de la voz. La experiencia de los autores en el uso de los rasgos SDC en el reconocimiento del locutor ha dado buenos resultados [44,45]. Estos rasgos se determinan a partir de cualquiera de los rasgos acústicos (ya sean LPCC, MFCC, etc.).

Dado un conjunto de rasgos $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T\}$ donde T es el número de tramas, y $\boldsymbol{x}_t = \{x_{t,0}, \ldots, x_{t,N-1}\}^T$ vector de rasgos N-dimensional, los rasgos SDC se calculan según la expresión [46]:

$$\Delta \boldsymbol{x}_{t,i} = \boldsymbol{x}_{t+iP+d} - \boldsymbol{x}_{t+iP-d}, \tag{14}$$

donde los parámetros que caracterizan al conjunto de rasgos N–d–P–k se pueden entender del esquema de la Fig. 3:

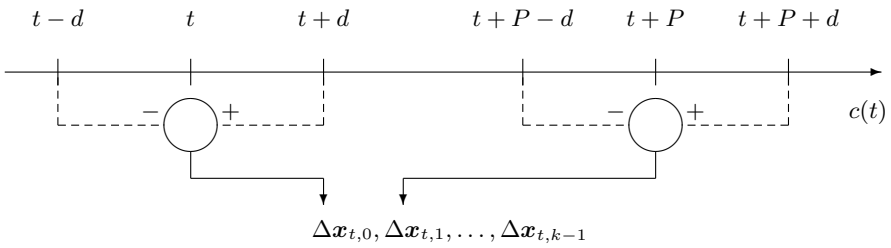


Figura 3. Representación esquemática para la extracción de los rasgos SDC

Nótese cómo la dimensionalidad aumenta de N para los rasgos originales, a kN para los SDC:

$$SDC[t] = \{\Delta \boldsymbol{x}_{t,0}^T, \ldots, \Delta \boldsymbol{x}_{t,i}^T, \ldots, \Delta \boldsymbol{x}_{t,k-1}^T\}. \tag{15}$$

Por tanto se impone una selección minuciosa de los valores mínimos de los parámetros N y k y de aquellos rasgos que permitan no sólo reducir el volumen de cómputo, sino también la construcción de un modelo más robusto.

El uso de una expresión similar a (13) reporta mejores resultados en la literatura [47]:

$$\Delta \boldsymbol{x}_{t,i} = \frac{\sum_{d=-D}^{D} d\, \boldsymbol{x}_{t+iP+d}}{\sum_{d=-D}^{D} d^2}. \tag{16}$$

2.6. Rasgos *wavelet*

En la ultima década el análisis *wavelet* ha demostrado ser efectivo para una variedad de problemas [48]. En particular, en esquemas de extracción de rasgos diseñados con el propósito del reconocimiento del habla, los wavelets han tenido dos salidas: la primera es el uso de la transformada *wavelet* como un decorrelador efectivo, en lugar de la DCT para la obtención de rasgos cepstrales [49], y la segunda es la aplicación directa de la transformada *wavelet* a la señal de voz, tomando aquellos coeficientes wavelet de mayor energía como rasgos [50], o utilizando las bandas de energía en lugar de las sub-bandas Mel [51]. En particular, en el área del reconocimiento del habla, la transformada *wavelet packet* (WPT), empleada para computar el espectro, fue propuesta en [52] y después fue usada en [51,53] y [54] para la construcción de los rasgos del habla, muy similares a la aproximación *Mel*. Se reportan recientes resultados [55] en el reconocimiento del lenguaje, como alternativa a los métodos fono-tácticos. En el campo del reconocimiento del locutor, recientes avances se han observado con la utilización de la transformada *wavelet* para la obtención de rasgos del locutor. Los trabajos del laboratorio de la Universidad de Patras utilizando WPT para extracción de rasgos de la voz en sub-bandas [48,56], han dado resultados alentadores evaluados en eventos NIST; por otra parte, [57] y [58] han desarrollado el análisis *wavelet* de rasgos suprasegmentales como el *pitch*, mientras [59] combina los coeficientes MFCC con coeficientes obtenidos de residuos *wavelets* por octavas (WOCOR, de sus siglas en inglés), los que pretenden representar las características espectro-temporales de la excitación de la fuente vocal.

La teoría *wavelet* brinda una estructura muy flexible para obtener representaciones de señales con buenas resoluciones en ambos dominios: frecuencial y temporal [60]. Permite también lidiar con problemas de ruido bien localizado en las frecuencias. La descomposición de una señal en un árbol *wavelet* se basa en la aplicación repetida de un par de filtros, uno pasa-bajo y otro pasa-alto, dando la posibilidad de dividir el espectro de frecuencia en intervalos de varios anchos de banda. En este análisis multirresolución de la señal, tanto la banda de bajas frecuencias como la de altas puede ser descompuesta resultando en una estructura de árbol binaria [61] mostrada en la Figura 4.

Cada nodo en el árbol tiene los subíndices (j, p), donde j es la profundidad y p es el número de nodos a la izquierda de este nodo en particular a la misma profundidad.

Los filtros mencionados anteriormente caracterizan la base ortogonal de $L^2(\mathbb{R})$ generada por la *wavelet* madre ψ de acuerdo al análisis multirresolución y satisfacen las condiciones

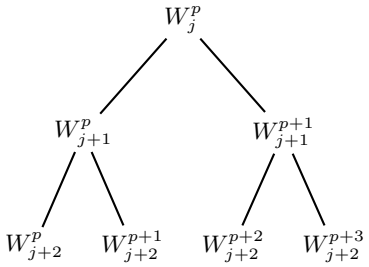


Figura 4. Estructura de árbol binaria del espacio *wavelet*

siguientes de ortogonalidad:

$$\begin{aligned} |G(\omega)|^2 + |G(\omega+\pi)|^2 &= 2, \\ G(\omega)H^*(\omega) + G(\omega+\pi)H^*(\omega+\pi) &= 0. \end{aligned} \tag{17}$$

donde H y G representan la transformada discreta de Fourier de los filtro pasa-bajo y pasa-alto respectivamente[3]. En este trabajo utilizamos las *wavelets* de Battle y Lemarié [62,63], la cual fue usada en [56] para un sistema de reconocimiento del locutor –obteniendo menor EER que los rasgos MFCC– y por los autores en [64], arrojando resultados de interés para la verificación e identificación del locutor. Estas *wavelets* son altamente localizadas en el tiempo debido a su caída exponencial y tienen un comportamiento similar a otras con un menor número de coeficientes. Su definición en el espacio de las frecuencias es:

$$\hat{\psi}(\omega) = \frac{\exp(-\frac{i\omega}{2})}{\omega^{m+1}} \sqrt{\frac{S_{2m+2}(\frac{\omega}{2}+\pi)}{S_{2m+2}(\omega) S_{2m+2}(\frac{\omega}{2})}}\,, \tag{18}$$

donde

$$S_n(\omega) = \sum_{k=-\infty}^{\infty} \frac{1}{(\omega+2k\pi)^n}. \tag{19}$$

2.7. Rasgos normalizados

La ecualización espectral es una técnica de normalización en el dominio de los rasgos que ha confirmado su efectividad en la reducción de los efectos lineales del canal y las variaciones espectrales a largo término [2,26]. El método de substracción de la media cepstral (*Cepstral Mean Subtraction*, CMS) es efectivo para aplicaciones de reconocimiento del locutor que utilicen expresiones suficientemente largas. Los coeficientes cepstrales son promediados sobre la duración de una expresión completa y el valor promedio es substraído de los coeficientes cepstrales en cada trama. La variación aditiva en el dominio del logaritmo del espectro puede compensarse bastante bien, sin embargo, como no puede evitarse que se remuevan algunos rasgos específicos del locutor, no es apropiada para aplicar en expresiones cortas [28].

[3] Una explicación más detallada del análisis wavelet se puede encontrar en [61].

Otra técnica de adaptación al canal es el filtrado cepstral paso-alto, que provee un nivel de robustez a bajo costo computacional [65,66]. En el método RASTA propuesto por [65], se aplica un filtro paso-alto o paso-banda a la representación logarítmica espectral del habla, modelando el mecanismo de la escucha humana que es más sensible a ciertas frecuencias y menos sensible a otras [67], las cuales son filtradas.

En el método de normalización CMN, el filtrado paso-alto se acompaña de la substracción del promedio a corto término de los vectores de coeficientes cepstrales: ambos algoritmos compensan los efectos del filtrado lineal porque fuerzan los valores promedios de los coeficientes cepstrales a cero, tanto en las muestras para el entrenamiento como para la clasificación.

Con el objetivo de reducir la influencia de las variaciones en las condiciones acústicas entre la fase de entrenamiento y de prueba, se ha propuesto en la literatura un método robusto de normalización para la reducción del ruido y de variaciones producidas por el canal, la normalización cepstral de media y varianza (*Cepstral Mean and Variance Normalization*, CMVN [68]). Este método normaliza la distribución espectral de las expresiones y reduce la variabilidad espectral intra-locutor de largo término. Asumiendo distribuciones gaussianas, CMVN normaliza cada componente del vector de rasgos según la expresión:

$$\hat{x}_{in} = \frac{x_{in} - \mu_i}{\sigma_i}, \tag{20}$$

donde x_{in} y $\hat{x}_{in}$ representan la i-ésima componente del vector de rasgos en la trama de tiempo n antes y después de la normalización respectivamente y, μ_i y σ_i son los estimados para la media y la varianza de la secuencia x_{in}.

Estos métodos han sido utilizados con regularidad en el reconocimiento del locutor [36].

2.8. Experimentos con LPCC y MFCC

A continuación se describen un grupo de experimentos realizados por los autores con rasgos LPCC y MFCC con el fin de evaluar sus desempeños usando diferentes métodos de normalización ante la diferencia de sesión y de canal, entre el entrenamiento y la prueba.

Todos los cálculos se realizaron en el programa Matlab. Se utilizaron grabaciones de 100 locutores de la base española *Ahumada* [69], las cuáles fueron recogidas en diferentes sesiones usando canales telefónicos y microfónicos. En este trabajo se usaron específicamente grabaciones cortas (de aproximadamente 5 segundos) de secuencias de dígitos y frases balanceadas.

Un primer experimento consistió en grabaciones de secuencias de dígitos y probar con grabaciones de secuencias de dígitos, y un segundo se basó en frases balanceadas fonéticamente. Ambos fueron realizados en 3 condiciones diferentes:

- Sesión variable y canal fijo. Aquí se usaron señales grabadas por el mismo micrófono pero en días diferentes.
- Sesión fija y canal variable. Aquí se usaron señales que fueron grabadas simultáneamente por un micrófono y un teléfono. O sea, solo hay variabilidad del canal.
- Sesión y canal variables. En este caso el entrenamiento se hizo con muestras microfónicas y las pruebas con señales telefónicas grabadas en días distintos.

En cada caso, se evaluaron 6 conjuntos de rasgos, etiqueteados de la siguiente forma:

- LPCC (18): los primeros 18 rasgos LPCC.
- (LPCC+Δ)-CMVN (36): a estos 18 rasgos LPCC se le calculan los *delta*-, y el conjunto de 36 rasgos es normalizado según CMVN.
- MFCC(Fisher)-CMN (12): se seleccionan los 12 mejores rasgos –de un conjunto inicial de 18 rasgos– según el criterio F-Ratio y se le aplica CMN.
- (MFCC(Fisher)+Δ)-CMVN (24): a los 12 mejores rasgos según F-Ratio se le calculan los *delta*-, y el conjunto de 24 rasgos es normalizado según CMVN.
- SDC(MFCC(1-12)-CMN) (60): a los primeros $N = 12$ rasgos MFCC se le aplica CMN y se les calculan los $k \cdot N = 60$ SDC.
- SDC(MFCC(1-12))-CMVN (60): a los primeros $N = 12$ rasgos MFCC se les calculan los $k \cdot N = 60$ SDC y luego son normalizados según CMVN.

En los cuadros 1–4 se resumen estos experimentos, reflejando el número de locutores clasificados correctamente en cada caso.

Rasgos	Números	Frases
LPCC (18)	57	71
(LPCC+Δ)-Norm (36)	95	98
MFCC(Fisher)-CMN (12)	92	94
(MFCC(Fisher)+Δ)-Norm (24)	96	99
SDC(MFCC(1-12)-CMN) (60)	83	99
SDC(MFCC(1-12))-Norm (60)	99	100

Cuadro 1. Sesión variable y canal fijo

Rasgos	Números	Frases
LPCC (18)	28	29
(LPCC+Δ)-Norm (36)	100	100
MFCC(Fisher)-CMN (12)	88	85
(MFCC(Fisher)+Δ)-Norm (24)	100	97
SDC(MFCC(1-12)-CMN) (60)	98	100
SDC(MFCC(1-12))-Norm (60)	100	100

Cuadro 2. Sesión fija y canal variable

2.8.1. Discusión de los resultados

En los casos donde se calculan los rasgos *delta*, se realizaron además experimentos sin normalizar los resultados. La clasificación fue mucho mejor con la aplicación de esta técnica.

Rasgos	Números	Frases
LPCC (18)	6	4
(LPCC+Δ)-Norm (36)	81	82
MFCC-CMN (18)	54	70
MFCC+Δ-Norm (36)	66	87
SDC(MFCC(1-12)-CMN) (60)	67	73
SDC(MFCC(1-12))-Norm (60)	76	83

Cuadro 3. Sesión y canal variables

Rasgos	Clasificación (%)
LPCC	32.5
(LPCC+Δ)-Norm	92.6
MFCC-CMN	80.5
MFCC+Δ-Norm	90.8
SDC(MFCC(1-12)-CMN)	86.7
SDC(MFCC(1-12))-Norm	93.0

Cuadro 4. Clasificación promedio de cada variante

En la mayoría de los experimentos realizados, los mejores resultados se obtuvieron con los SDC. Estos tienen la desventaja de que usan un vector de rasgos mucho más grande que el resto de las variantes analizadas, lo cual aumenta el volumen de cálculo. No obstante, si aumentamos el número de rasgos usados en el caso de los LPCC o los MFCC, la clasificación no mejora considerablemente –y hasta puede empeorar al enfrentarse al problema de la dimensionalidad–, ya que en estos casos, a medida que el índice de los rasgos aumenta, la información que aportan estos para la identificación del locutor disminuye.

El clasificador es otro factor determinante en la identificación del locutor. Como no era objetivo en esta parte del trabajo, nos hemos limitado a usar un cuantificador vectorial (VQ) para crear los modelos y una medida de distorsión para hallar la similaridad. Existe un gran número de técnicas como los modelos de mezclas gaussianas (*Gaussian Mixture Models*, GMM), las máquinas de vectores de soporte (*Support Vector Machines*, SVM), técnicas de fusión de clasificadores y otros que serán analizadas en detalle en el Capítulo 4.

2.9. Reducción de la variabilidad

Un factor que afecta sensiblemente el rendimiento de los sistemas de RAL, es la diferencia entre las condiciones acústicas en que se realiza el entrenamiento y en las que tiene lugar la etapa de prueba. A esta diferencia contribuyen grandemente los cambios de canal y de ambiente. Las fallas de estos sistemas no se deben solamente a las variaciones de canal, ruido ambiente o similaridades entre locutores diferentes. También influyen las diferencias entre dos locuciones hechas por la misma persona, conicido como variabilidad intralocutor.

El problema que, en general, engloba las dificultades anteriores, es conocido como variabilidad de sesión y se refleja cuando un modelo entrenado bajo un conjunto de condiciones es empleado para probar datos obtenidos bajo condiciones muy distintas.

Actualmente existen diversas y muy variadas técnicas aplicadas a la compensación o eliminación de la variabilidad de canal; CMS y RASTA –mencionadas anteriormente– mejoran el rendimiento de los sistemas de RAL en presencia de distorsiones convolucionales y ruido aditivo.

Otra propuesta interesante para contrarrestar las distorsiones del canal es la conocida por *Feature Warping* [70,71].

El Mapeo de rasgos (*Feature Mapping*) es otra aproximación, que explota la información *a priori* de un conjunto de modelos entrenados bajo condiciones controladas, con el objetivo de mapear los vectores de rasgos a un espacio de rasgos independiente del canal. El inconveniente que tiene esta aproximación es que requiere que los datos de entrenamiento estén etiquetados para identificar las condiciones que se quieren compensar, no obstante, en [72] se propone una técnica de mapeo de rasgos para lidiar con este inconveniente.

Igualmente se han propuesto técnicas, que actúan a nivel de los modelos, para contrarrestar la variabilidad de sesión. Estas han demostrado que son capaces de compensar las variaciones sin necesidad de explicitar con etiquetas las condiciones en las que fueron obtenidos los datos, además la mayoría de estas técnicas se basa en modelar la variabilidad de las frases restringiéndolas a un espacio de menor dimensión.

En [73,74] se plantea que la variabilidad de sesión es una importante fuente de error en el reconocimiento del locutor. Aquí se propuso un modelo basado en el análisis de factores (FA) y muchos investigadores han confirmado su efectividad para sistemas basados en GMM [75]. FA asume que las variables pueden ser agrupadas por sus correlaciones, aquellas que pertenezcan a un mismo grupo estarán altamente correlacionadas entre si y tendrán poca correlación con las variables de otros grupos. Se dice entonces que cada grupo de variables, representada por un factor, es producto de las correlaciones entre las observaciones. Este es un método estadístico usado para explicar la variabilidad entre variables observables en función de variables no observables llamadas factores, las variables observables se modelan como combinaciones lineales de los factores más un término de error.

Otras importantes técnicas de compensación a nivel de modelo son: NAP (*Nuisance Attribute Projection*) [76,77] y WCCN (*Within-Class Covariance Normalization*) [78] para SVMs.

2.10. Detección de actividad de voz

Detección de actividad de voz (VAD, por sus iniciales en Inglés) es el algoritmo usado en procesamiento del habla para detectar la presencia o ausencia de voz humana en una señal de audio. Los VAD pueden no solo indicar la presencia o ausencia de voz, también pueden indicar voz sonora o no sonora.

La voz varía debido a las diferencias introducidas por micrófonos, teléfonos, género, edad y otros factores; pero la principal causa son los ruidos en la señal, los cuales pueden provocar un mal funcionamiento en los algoritmos de los sistemas de procesamiento del habla cuando trabajan bajo condiciones extremas de ruido. Comunicaciones inalámbricas, ayuda a la audición digital o reconocimiento robusto de voz y sistemas de reconocimiento automático

del locutor, son algunos ejemplos de tales sistemas que frecuentemente requieren técnicas de reducción de ruido combinadas con un algoritmo robusto VAD.

Los algoritmos VAD existentes utilizan para la detección de voz y no voz rasgos como: medida de distancia LPC Itakura, niveles de energía, frecuencia fundamental, cruces por cero, modelos adaptables de ruido, etc. [79,80,81,82].

3. Principales métodos de selección de rasgos del habla

Una posible manera de elevar la efectividad de un sistema de reconocimiento de patrones es elevar la dimensionalidad de los rasgos. Sin embargo, esto causa más problemas que ventajas. El problema conocido como la *maldición* de la dimensionalidad (*curse of dimensionality*) [83] consiste en que la cantidad de datos que se requieren para lograr una representación adecuada del patrón, crece exponencialmente con la dimensión de los rasgos, lo que afecta la eficiencia del reconocedor, pues incrementa el costo computacional y los requerimientos de memoria al necesitar más datos para entrenar al reconocedor. Si no se cuenta con suficientes datos, el incremento de la dimensionalidad puede empeorar la efectividad del reconocedor en lugar de mejorarla [84]. Se ha demostrado que la efectividad de un reconocedor mejora para algunos tipos de rasgos al reducir la dimensionalidad de los mismos [85]. Para solucionar dicho problema, se requiere reducir adecuadamente la dimensionalidad del espacio de rasgos, obteniendo una compactación del mismo y elevando la eficiencia del reconocedor que trabajará con un conjunto menor de rasgos, menos redundantes, más relevantes y discriminativos. En el reconocimiento del locutor, se trata de seleccionar del conjunto de K rasgos acústicos extraídos, aquellos que sean más robustos ante la variabilidad del locutor y del canal, que maximicen la independencia entre ellos y que minimicen el error de clasificación, lo cual no es fácil de obtener de forma integral [86]. Es muy común utilizar para la clasificación en reconocimiento del habla y del locutor un vector de 12 coeficientes cepstrales, 12 rasgos *delta* y 12 rasgos *delta-delta*, con una dimensionalidad de $K = 36$ [40,43]; sin embargo, se ha comprobado que hay rasgos redundantes y de baja relevancia [87,88], que no aportan en igual medida al reconocimiento del locutor. Un método de selección de rasgos que minimiza el error de clasificación, consiste en escoger el mejor sub-grupo de k rasgos de todos los K rasgos que representan al patrón, que brinde mejor probabilidad de clasificación correcta. Existen diversos métodos de búsqueda para la selección de rasgos minimizando el error de clasificación [86,89]:

1. Búsqueda exhaustiva(*Exhaustive search*): es el método más completo de búsqueda para selección de rasgos, se evalúa el error de clasificación con todas las posibles combinaciones de sub-grupos de k rasgos, requiere un enorme esfuerzo computacional, y no se utiliza en la práctica.
2. Método de k mejores (*K-best Method*): es el método más simple, el mejor subgrupo de rasgos esta compuesto por los k mejores rasgos, considerados independientes. Sin embargo, un subgrupo de los k mejores rasgos individuales no necesariamente es el mejor subgrupo de rasgos. Un ejemplo típico es la utilización de la razón de Fisher (*F-ratio*) para obtener aquellos rasgos acústicos que exhiban alta variabilidad inter-locutor y baja variabilidad intra-locutor de forma individual , brindando los mayores *F-ratio*.

Sin embargo, se requiere evaluar el reconocedor con muchos subgrupos de rasgos, pues se ha comprobado que los sub-grupos de k rasgos acústicos que mejor *F-ratio* poseen, no necesariamente brindan los menores errores de clasificación [29].

3. Selección hacia delante (*Forward Selection*): Este método [90] comienza con el espacio de rasgos vacío y se van adicionando rasgos iterativamente, la prueba inicial se hace con cada rasgo individual, uno a uno, seleccionando los mejores k rasgos simples. Después se prueba con dos rasgos, incluyendo uno de los mejores seleccionados previamente y cada uno de los restantes $K - k$ rasgos, el ciclo se repite hasta que se seleccionen los rasgos que se deseen.
4. Selección hacia atrás (*Backward Selection*): Este método es una técnica de búsqueda paso a paso, llamada estrategia de *knock-out* [90,91] y comienza con el espacio total de K rasgos, todos los K sub-grupos de $K-1$ rasgos se usan para calcular el comportamiento y determinar el mejor sub-grupo de $K - 1$ rasgos, el rasgo no usado queda fuera. El proceso se repite con $K-1$ sub-grupos de $K-2$ rasgos hasta que se llegue al sub-grupo de k rasgos que se desean.
5. Algoritmo $l-r$ ($l-r$ *algorithm*): El algoritmo $l-r$ [92] usa la selección hacia delante y hacia atrás par obtener el mejor comportamiento del procedimiento de selección. Para cada iteración el algoritmo usa el procedimiento hacia delante para agregar l rasgos al sub-grupo y usa el procedimiento hacia atrás para eliminar los peores r rasgos del sub-grupo, hasta lograr los k rasgos deseados. Existe una variante dinámica del algoritmo $l - r$ conocida como *Sequential Floating Forward Sequence* (SFFS) [93], que consiste en aplicar, después de cada paso hacia adelante, un número de pasos hacia atrás hasta que el sub-grupo resultante sea mejor que los anteriores, no habrá pasos hacia atrás si no se mejora el comportamiento.
6. Programación dinámica (*Dynamic Programming*): se utiliza para obtener el número óptimo de rasgos con mucho menos cálculos que la búsqueda exhaustiva, se trata de una técnica de optimización que cuando se aplica a la selección de rasgos en conjunto con una ecuación funcional, permite la selección de rasgos que tienen la máxima efectividad [94].

Otros métodos de selección asignan a los rasgos una puntuación y seleccionan los rasgos mejor puntuados para hacer la clasificación, el resultado depende mucho del método de puntuación que se utilice [84]. La puntuación de los rasgos se obtiene por medio de una transformación lineal a los K rasgos originales, obteniendo otros k rasgos ($k < K$) que preservan mucha de la información de los rasgos originales, logrando reducirse la dimensionalidad [29].

En el caso del habla, los rasgos mas usados, como los cepstrales en escala *Mel* (MFCC) y los pares de líneas espectrales (LSP), son bastante no-correlacionados [10]. No obstante, se aplican métodos de transformación a dichos rasgos u otros, para obtener un nuevo espacio de rasgos, más discriminativos y aún menos correlacionados. Estos métodos incluyen, entre otros, el análisis de discriminante lineal (*Linear Discriminant Analysis, LDA*) [83,95], el análisis de componentes principales (*Principal Component Analysis, PCA*) o como también se le conoce, transformada de Karhunen-Loève (KLV) [95] y el análisis de componentes independientes (*Independent Component Analysis, ICA*) [96].

Estos tres métodos se formulan como una transformación lineal que proyecta los vectores de rasgos en un sub-espacio transformado, definido por direcciones relevantes. La matriz de la transformación A se estima de forma tal que, desde el punto de vista de la clasificación, se reduzca la redundancia y se retenga la información relevante. Esto debe optimizar el comportamiento del vector transformado. Al remover los rasgos perjudiciales y confusos, con gran probabilidad, debe también mejorarse y robustecerse el estimado de los parámétros del modelo.

3.1. Análisis de componentes principales

Este método de análisis estadístico multivariado [95,97] consiste en computar los vectores característicos de la matriz de covarianza $\frac{1}{n-1}XX^T$, organizarlos de acuerdo a sus valores propios en orden descendente y, finalmente, construir la matriz de proyección A (*Karhunen-Loève Transform*, KLT) con los K mayores vectores característicos en relación con su varianza. Cada vector $\boldsymbol{x}$ se preprocesa de acuerdo a la expresión $\boldsymbol{y} = A(\boldsymbol{x} - \boldsymbol{\mu})$, donde $\boldsymbol{\mu}$ representa el vector de rasgos promedio. La transformación KLT de-correlaciona los rasgos y brinda el más pequeño error de reconstrucción posible entre todas las transformaciones lineales, o sea, el menor error medio cuadrático (MSE, de sus siglas en inglés) entre los vectores de datos en el espacio original de D rasgos y los vectores de datos en el espacio proyectado de K rasgos. Este método puede usarse para de-correlacionar combinaciones de rasgos estáticos y dinámicos MFCC, PLP y otros [28]. Desafortunadamente, PCA no minimiza el error de clasificación [87], o sea, no es necesariamente óptimo para discriminar entre clases [98]. Como el reconocimiento del locutor es un problema discriminativo, generalmente se aplican otros métodos de transformación para reducir la dimensionalidad de los vectores de rasgos [29].

3.2. Análisis discriminante lineal

Conocido también como *Discriminante Lineal de Fisher*, se propone hallar la matriz de transformación A que maximice el criterio de separabilidad entre clases [83]. Esto se hace computando las matrices de varianza intra-clase W y entre-clase B, hallando los vectores característicos de $W^{-1}B$, organizándolos de acuerdo a sus valores propios en orden descendente, y finalmente construyendo la matriz de proyección A con los K mayores vectores característicos, que definen los K más significativos hiperplanos. LDA (de sus siglas en inglés) asume que las clases son linealmente separables, que todas poseen una distribución gaussiana simple y comparten una covarianza común intra-clase. Adicionalmente, como método de entrenamiento supervisado, requiere el etiquetado de las muestras con la identificación de cada clase. LDA es usualmente aplicado sin asumir un modelo específico para las clases, brindando un comportamiento razonable en muchas aplicaciones [99], dado por la estabilidad de las estadísticas entre-clases e intra-clases [100]. LDA ha sido utilizado en reconocimiento del locutor en ambientes ruidosos [35,101,102,103,104], con buenos resultados. En [105] se reportan resultados con análisis de discriminante no lineal.

3.3. Análisis de componentes independientes

Es una técnica moderna que intenta reducir la redundancia en el espacio de rasgos originales [96]. Mientras que PCA remueve las dependencias de segundo orden, ICA remueve también las de orden superior, minimizando aún más la información mutua entre rasgos, proyectándolos en la dirección de máxima independencia. Mientras PCA representa datos con distribuciones gaussianas, ICA puede descomponer datos con una naturaleza estadística esparcida, no gaussiana. De hecho, ICA fue diseñado originalmente para resolver el problema de la separación a ciegas de la fuente u origen (*blind source separation, BSS*). Las señales observadas son asumidas como una combinación lineal de varias señales de procedencia, no-gaussianas y estadísticamente independientes. La tarea de ICA es recuperar las señales de procedencia a partir de las señales observadas, o sea, buscar aquella dirección que sea mejor para separar las fuentes. Los rasgos extraídos del habla en ambientes reales presentan componentes con distribuciones desconocidas, incluso no gaussianas, producto del ruido ambiente, de un locutor no deseado o de música en el efecto *cocktail party*. La descomposición y recuperación de dichos componentes en los rasgos del habla para sistemas robustos de reconocimiento del locutor, se enfrentan con ICA en lugar de PCA por su mayor capacidad de seleccionar los rasgos más representativos. No obstante, presenta como inconveniente que requiere de una segunda fuente de voz para poder aplicar BSS [106].

3.4. Algoritmos genéticos

Los algoritmos genéticos fueron propuestos por [107], son técnicas de búsqueda heurística basadas en estrategias de evolución biológica y se utilizan en varias disciplinas como un nuevo medio de optimización de sistemas complejos. La idea básica de los *GA* (por sus siglas en inglés) es la de *selección natural* el principio de que *sobrevivan los más aptos* aplicando tres operaciones básicas: selección de la funcion de aptitud *fittest*, cruce y mutación. Los *GA* son capaces de optimizar un sistema complejo sin necesidad de conocer sobre él, lo que posibilita una retroalimentación entre la salida de un sistema de reconocimiento y el extractor de rasgos a optimizar. En [108], se utiliza los *GA* para el diseño de nuevos extractores de rasgos acústicos y en [87,88] se utiliza para determinar pesos para la selección de rasgos acústicos.

3.5. Teoría de la información

La teoría de la información [109] es una herramienta poderosa que ha demostrado recientemente ser muy útil para el análisis de la selección de rasgos en el reconocimiento del locutor. La *información mutua* (MI, por sus siglas en inglés) entre locutores y rasgos se utiliza para la selección de éstos. La mayor ventaja del uso de *MI* como medida de comportamiento, en lugar de utilizar la probabilidad de error de clasificación, es que ésta puede computarse analíticamente y maximizarse mejor que el error de clasificación. En los trabajos de [110,111,112], la transformación lineal o no lineal de un espacio de rasgos acústicos de gran dimensión a uno de baja dimensión, se optimiza maximizando la *MI*. En [113,114], se utiliza la *MI* para la selección o combinación de rasgos acústicos. En [48],

se evalúan varios métodos para selección y puntuación de rasgos, basados en su entropía, propuestos por [115]: evaluación de rasgos por ganancia de información, razón de ganancia e incertidumbre simétrica.

A continuación se enuncian los conceptos más importantes relacionados con este tema.

3.5.1. Entropía

Se define –según Shannon– la *entropía* de un conjunto discreto X como:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x), \tag{21}$$

donde $p(x)$ es la probabilidad de ocurrencia del elemento x ($\sum_x p(x) = 1$). Esta es una medida del grado de desorden o de desconocimiento de un sistema. La entropía así definida es no negativa. La unidad en que se mide depende de la base del logaritmo utilizada, en nuestro caso dicha base es 2, entonces medimos la entropía en bits.

Para un sistema completamente determinado en un estado x_0, $H(X) = 0$ (pues tenemos que $\forall x \in X, p(x) = \delta_{xx_0}$). El máximo de (21) se obtiene cuando $p(x) = \Omega_X^{-1}, \forall x \in X$, donde Ω_X es el número de elementos en X. Es decir, la distribución uniforme es la configuración más desordenada para un conjunto discreto; el conocimiento de los estados en este sistema es el mínimo posible.

La *entropía conjunta* de X e Y se define de forma análoga:

$$H(X,Y) = -\sum_{x,y} p(x,y) log_2 p(x,y). \tag{22}$$

La *entropía condicional* es la información adicional que aporta Y conocida X:

$$H(Y|X = x) = -\sum_{x,y} p(y|x) log_2 p(y|x)$$

$$\begin{aligned} H(Y|X) &= \sum_{x,y} p(x) H(Y|X = x) = -\sum_{x,y} p(x,y) log_2 \frac{p(x,y)}{p(x)} \\ &= -\sum_{x,y} p(x,y) log_2 p(x,y) + \sum_{x,y} p(x,y) log_2 p(x) \\ &= H(X,Y) - H(X). \end{aligned} \tag{23}$$

3.5.2. Información mutua

Introducimos ahora el concepto de información mutua, la cual es una medida de la cantidad de información que una variable aleatoria comparte de otra.

Se define la información mutua entre dos conjuntos X e Y como:

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{24}$$

que mide la información común que es común a ambos conjuntos. Esta expresión es simétrica respecto a X e Y. Usando (22) y (23) se obtienen definiciones análogas:

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{25}$$

$$I(X,Y) = \sum_{x,y} p(x,y) log_2 \frac{p(x|y)}{p(x)} = \sum_{x,y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}. \tag{26}$$

3.5.3. Propiedades

A continuación se resumen las propiedades básicas relacionadas con los conceptos introducidos en este capítulo:

- $I(X,Y) \geq 0$,
- $H(X,Y) \geq H(X) \geq H(X|Y)$,
- $I(X,Y) \leq H(X)$, solo igual si $H(X|Y) = 0$ (X completamente determinado por Y),
- si X y Y son independientes estadísticamente,

$$H(X|Y) = H(X) \quad \text{e} \quad I(X,Y) = 0,$$

- $H(X|X) = 0$, luego $I(X,X) = H(X)$. Es decir, $I(X,X) \geq I(X,Y)$, una variable contiene más información de ella misma que la que pueda proveer cualquier otra variable.

De acuerdo a las definiciones dadas en esta sección, los rasgos con menor información del locutor tienen un valor menor de información mutua con la clase de los locutores. Los mejores K rasgos – siguiendo este criterio – son aquellos $\{y_{i_1}, \ldots, y_{i_K}\} \subset \{y_1, \ldots, y_N\}$ que maximicen la información mutua con la clase de los locutores:

$$\{y_{i_1}, \ldots, y_{i_K}\} = \underset{\{y_{j_1},\ldots,y_{j_K}\}}{\arg\max}\ I(\{y_{j_1}, \ldots, y_{j_K}\}, \mathcal{S}), \tag{27}$$

donde N es el número total de rasgos.

4. Métodos de clasificación

A partir de los rasgos acústicos del habla de cada locutor, previamente extraídos y seleccionados, se requiere encontrar un modelo que clasifique efectivamente al mismo y sea lo suficientemente robusto ante la variabilidad del habla.

Los métodos de clasificación han evolucionado en el tiempo. En la décadas del 70 predominó la clasificación comparando plantillas de palabras, en los 80 se clasificó aplicando la distorsión dinámica en el tiempo (DTW, por sus siglas en inglés) y la cuantización vectorial (VQ). A partir de mediados de los 90 se desarrollan enfoques estadísticos de clasificación como los HMM y los modelos de mezclas gaussianas (GMM), lográndose los mejores resultados de clasificación hasta el momento.

En el 2000 comienzan a aplicarse diferentes formas de adaptar o combinar los modelos GMM de los locutores con los modelos universales de background (UBM, por sus siglas en

inglés), predominando la adaptación máximo *a posteriori* (MAP), que ha sido clave en las mejoras de los clasificadores en esos años, alcanzando el estado del arte de dichos modelos durante la evaluación del *National Institute for Standards and Technology* (NIST, 2004).

En este momento se comienza a desarrollar un enfoque de clasificación discriminativa a partir de los modelos GMM, como los GMM Supervector Linear Kernel (GSL) que utilizan supervectores construidos de las medias de los modelos GMM adaptados MAP, para clasificar los locutores con máquinas de soportes vectoriales (SVM).

4.1. Métodos de clasificación de rasgos acústicos

- **Comparación de patrones**: ajuste dinámico en el tiempo **DTW** que alinea las secuencias de rasgos de entrenamiento y prueba para su comparación.
- **Vecino más cercano**: cuantificación vectorial **VQ** para cada vector de rasgos de la expresión de prueba, se mide su distancia a los vectores de entrenamiento más cercanos.
- **Redes neuronales**: como el perceptrón multi-capa (MLP, por sus siglas en inglés), la función de base radial (RBF, por sus siglas en inglés) o las redes neuronales en árbol (NTN, por sus siglas en inglés), que son entrenadas para discriminar entre un locutor y algunos locutores alternativos.
- **Modelos de Markov**: como los HMM y los GMM, que reflejan de forma estadística cómo se expresa cada locutor.

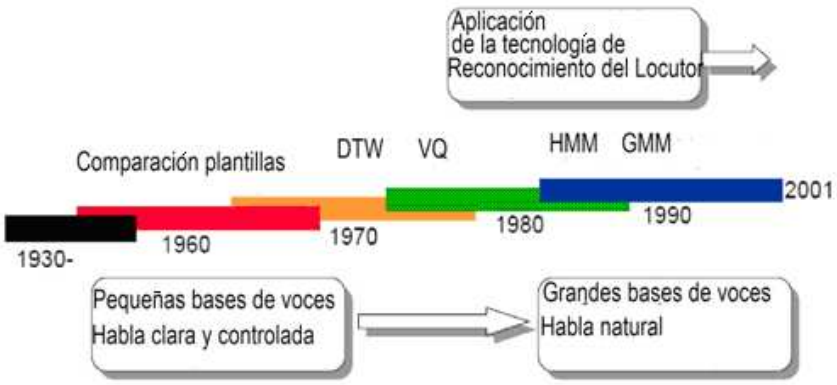


Figura 5. Evolución de la utilización de herramientas de Reconocimiento de Patrones en el reconocimiento automatizado del locutor hasta el año 2001

4.2. Descripción de los métodos

4.2.1. Distorsión dinámica en el tiempo

Una descripción del método de determinación de distancias entre dos series de tiempo para la comparación de plantillas de habla (*Dynamic Time Warping*) DTW, puede encontrarse en [40] y [23]. Esta técnica fue utilizada inicialmente en reconocimiento de palabras [116] y ha sido usada desde los años 80 en sistemas de reconocimiento del locutor dependiente del texto, comparando diferentes realizaciones temporales de las mismas expresiones. En este sentido, la mayor puntuación de similaridad o menor distancia se obtiene para diferentes realizaciones de un mismo *password* de un mismo locutor.

El algoritmo DTW busca el mejor trayecto de alineamiento por medio de técnicas de programación dinámica entre la expresión de entrada y la referencia, realizando un mapeo lineal segmentado entre uno o ambos ejes de tiempo para alinear las dos señales; al concluir el mapeo, la distancia acumulada entre las dos expresiones es la base de la puntuación [29]. Si las expresiones son idénticas en el tiempo, el trayecto de alineamiento es una diagonal; si no son idénticas, las desviaciones de la diagonal representan las distancias requeridas a distorsionar.

El método puede aplicarse también para alinear las variaciones en el tiempo de rasgos correspondientes a la configuración dinámica de los articuladores y el tracto vocal, como la energía [29] y los coeficientes LPC y dinámicos [8].

Hay dos factores que afectan el funcionamiento del DTW en sistemas dependientes del texto: la detección del punto final y el establecimiento de restricciones al trayecto de alineamiento local. Debido a la simplicidad del sistema es fácilmente aplicable en tareas como control de acceso con *password*. Sin embargo, no puede usarse en aplicaciones donde se induzca el *password*, porque se requeriría que el sistema tuviera las plantillas de referencia de todos los posibles *passwords* para cada locutor autorizado. Además el método es altamente dependiente de las expresiones de referencia, no permitiendo variabilidad en la señal de voz.

4.2.2. Métodos de cuantización vectorial

En la cuantización vectorial cada vector N-dimensional de entrada (un punto en el espacio N-dimensional) se representa por el más cercano *codevector* o centroide de un pequeño grupo de vectores o *codebook* altamente representativos de la distribución de vectores de entrada en el espacio N-dimensional. Este *codebook* se selecciona con los mejores representantes de los diferentes *clusters* o grupos en los que se hayan dividido los datos de entrada.

El reconocimiento del locutor basado en el modelado de múltiples plantillas para representar las expresiones de voz utiliza el método VQ [8][29][117]; los modelos del locutor se entrenan agrupando los vectores de rasgos de cada locutor en grupos no solapados. Cada grupo estará representado por su centroide (o vector medio). El conjunto de centroides de cada locutor es el *codebook* que representa al locutor. El tamaño de cada *codebook* es significativamente menor que el conjunto de entrenamiento, pero sus centroides siguen la

misma distribución que los vectores de entrenamiento, reduciendo el tamaño de los datos y preservando la información de la distribución original.

Existen dos criterios de diseño en la generación del *codebook*: el método empleado y su tamaño. La experiencia demuestra que el aumento del tamaño del *codebook* reduce los errores de reconocimiento [117][10], sin embargo, si el *codebook* es muy grande se entrena a los vectores de entrada pero no a la distribución quedando el modelo sobreajustado (*overfitting*), según Kinnunen [10] la pretensión de que el mejor modelo del locutor es el mismo conjunto de datos en si, no es verdadera en general.

En cuanto al método de creación del *codebook*, existen dos algoritmos de entrenamiento [10], supervisados o no. En los no supervisados cada *codebook* de cada locutor se entrena independientemente, mientras que en el caso de los supervisados las relaciones entre los *codebooks* se tienen en cuenta para reducir los posibles solapamientos, usualmente se utilizan los métodos no supervisados porque no requieren control del entrenamiento.

El más popular y simple de los métodos de entrenamiento no supervisados es el conocido como *Generalized Lloyd Algorithm* (GLA) o algoritmo de Linde-Buzo-Gray (LBG), por sus inventores [118] que tiende a reducir los errores de cuantización.

Un método de entrenamiento discriminativo supervisado es el *group vector quantization* (GVQ) propuesto en [119], el cual entrena al *codebook* individualmente y después se realiza un ajuste fino al entrenamiento para enfatizar las diferencias inter-locutores.

Otro método de entrenamiento discriminativo, que tiende a minimizar la razón de error de clasificación es el conocido como *learning vector quantization* (LVQ), propuesto por Kohonen en [120].

Los elementos necesarios para un sistema de reconocimiento del locutor usando VQ son:

- *Datos del entrenamiento*: vectores de rasgos obtenidos de la parametrización de la base de locutores.
- *Algoritmo de agrupamiento*: permite dividir el espacio de datos del entrenamiento en *clusters*, los cuales serán representadas en el *codebook* con su respectivos centroides. El entrenamiento del *codebook* se realiza por medio del algoritmo GLA, por medio de un divisor binario u otro. Según Kinnunen [121], el algoritmo seleccionado no es vital en el comportamiento del reconocimiento.
- *Asignación de vectores*: determina el procedimiento de búsqueda del centroide más cercano para cada vector de rasgos de entrada de un locutor desconocido. Se realiza usualmente por medio de la búsqueda del vecino más cercano o por medio de búsqueda en árboles [122] en el caso del divisor binario.
- *Medida de similaridad*: es la medida básica de distorsión que posibilita llevar a cabo las asignaciones de los vectores de rasgos de entrada a los diferentes grupos o *clusters*. La medida de similaridad más extendida es la distancia euclidiana o euclidiana cuadrada, que tiene un sentido geométrico en el espacio N-dimensional de vectores característicos (puede demostrarse que la distancia euclidiana entre dos vectores cepstrales mide la distancia entre los correspondientes espectros logarítmicos a corto término [23]. En ocasiones, puede usarse la distancia de Manhattan, de Mahanalobis u otras [23,83]. Se han propuesto diversas modificaciones a la medida de la similaridad [121][123][124]. En [125], se lleva a cabo una comparación entre varias medidas de similaridad y se establece

que el entrenamiento discriminativo y la medida de distancia normalizada son los más eficientes y pueden combinarse.

Para la aplicación de VQ al reconocimiento de locutores, se construye un *codebook* por cada locutor de la base, en el caso de la identificación del locutor se computa la distorsión de los vectores de rasgos de entrada con respecto a cada *codebook* y la menor distorsión nos indicara el locutor identificado; en el caso de verificación del locutor, se computa la distorsión de los vectores de entrada con respecto al *codebook* del locutor que clama su identidad y se compara con un umbral, si la distorsión es menor que el umbral, se acepta como verificado el locutor. Gran cantidad de sistemas de reconocimiento del locutor utilizan este esquema general.

4.2.3. Métodos discriminativos: redes neuronales

Las redes neuronales (artificial neural networks, ANN) son modelos computacionales que emulan el comportamiento del cerebro por medio de topologías que reflejan la interconexión de las células nerviosas [126]. Una ANN consiste en un grupo de "neuronas" (nodos o celdas) que se distribuyen en capas y se interconectan por medio de caminos pesados. Cada neurona es un elemento procesador que entrega una salida simple a partir de múltiples entradas, dicha salida usualmente es controlada por una función de activación.

En sentido general, las ANN poseen una capa de entrada, una capa de salida y una o más capas ocultas, estando conectadas las salidas de una capa a las entradas de la próxima. Una ANN "se entrena" ajustando los pesos de los caminos de las uniones entre neuronas. El entrenamiento es supervisado si se conocen los patrones a aplicar a la capa de entrada y las clases correspondientes a obtener en la capa de salida, las neuronas correspondientes a las capas ocultas se entrenan de forma tal que cuando se cargue el patrón a la entrada, la salida de la clase a la cual corresponde el mismo, se active [127]. Estas redes también se conocen como maquinas de aprendizaje o "perceptrones". El entrenamiento es no supervisado si no se conocen las clases de salida, existen también métodos de entrenamiento híbrido [128].

Las ANN más usadas para clasificación en reconocimiento del locutor, y que han obtenido relativamente buenos resultados en tareas de moderada complejidad son [129]:

- *Multi-Layer Perceptron* (MLP) [130][131]: se trata de una red de varias capas, cuya función de activación de las neuronas es del tipo "sigmoide", cada salida corresponde a una clase. El entrenamiento supervisado se lleva a cabo de acuerdo a una función de costo, la forma más usual de entrenamiento es por medio del algoritmo de retropropagación (*back propagation*"). El MLP es robusto al ruido y permite tener en cuenta el contexto de la señal. Puede encontrase una información bastante detallada de los MLP en [128]. El uso de un solo MLP en identificación del locutor se limita cuando crece la población de locutores a identificar, al elevarse el tiempo de entrenamiento por cada locutor que se incorpore. Por esta razón, en [130] proponen el uso de una ANN por locutor, que se entrena para salida igual a uno para la voz del locutor a identificar y salida igual a cero para las voces de los restantes locutores, existiendo tantas ANN como locutores. La clase cero se construye de un cuerpo de locutores común a todas las ANN. En identificación, la voz de prueba se evalúa en todas las redes; en verificación,

solo se requiere evaluar una ANN. Otra alternativa propuesta por [132] para reducir el tiempo de entrenamiento, consiste en realizar una partición binaria de los locutores, cada ANN se compone de dos clases ocultas y se especializa en discriminar entre dos locutores. Con una población de N locutores, se requieren $N(N-1)/2$ redes ANN. Un locutor se identifica entre N locutores, después de $N-1$ decisiones binarias. Se comprobó que se reduce el tiempo de entrenamiento al incrementarse N, pero el número de ANN es proporcional N^2, lo cual es prohibitivo para grandes poblaciones de locutores.

- *Radial Basis Functions* (RBF) [133][134]: constituye una alternativa a MLP que reduce el tiempo de entrenamiento con las mismas propiedades discriminativas de MLP, es una red de dos capas con funciones gaussianas en las neuronas de la primera capa y funciones lineales en la capa de salida, teniendo un comportamiento similar al modelo GMM. Las RBF poseen un método híbrido de entrenamiento, donde se forma una base para el espacio de entrada, en que la activación de uno de los nodos es una función de su proximidad al patrón de entrada en cada momento. Estas redes han demostrado poseer mejor poder discriminativo que las MLP y la VQ.
- *Learning Vector Quantization* (LVQ) [135]: la arquitectura LVQ utiliza varias celdas de salida para cada clase, es similar a la VQ y su entrenamiento supervisado se basa en el principio de clasificación del vecino más cercano. Sin embargo, los vectores de referencia no se escogen de los patrones de entrada sino que se entrenan de acuerdo al objetivo de la clasificación, permitiendo identificar las fronteras de las clases con más precisión que la VQ con muchas menos referencias. Una aplicación de identificación utilizando LVQ dependiente del texto que brindó buenos resultados, fue hecha por Bennani, 1990 [135].

Otras configuraciones alternativas de ANN, que han sido aplicadas al reconocimiento del locutor, son:

- *Task Decomposition Neural Network* (TDNN): este tipo de ANN es una alternativa a los MLP para la reducción de los tiempos de entrenamiento para largas poblaciones utilizando el concepto de la arquitectura modular y brindando soluciones al incremento de la complejidad de los modelos. Una primera aplicación fue la de Hamsphire en 1990, que propone una arquitectura para el reconocimiento del locutor a partir del reconocimiento de fonemas. Posteriormente, Bennani [136] presentó un TDNN constituido por un detector de topología y varios módulos expertos. Los rasgos de entrenamiento de toda la población se dividen en subgrupos por medio de una técnica de clasificación k-means y los locutores se etiquetan por el número de los subgrupos donde hayan caído más de sus rasgos, el detector de topología se entrena para clasificar a los locutores de acuerdo al etiquetado, cada modulo experto esta formado por una TDNN de tres capas y se dedica a discriminar entre locutores de la misma topología.
- *Task Decomposition Neural Network* (TDNN) + *Hidden Markov Models* (HMM): para facilitar la incorporación de nuevos locutores, propone un sistema modular híbrido, que combina módulos TDNN con un modulo final HMM, el método ha sido aplicado también por Bennani en 1992 y 1994, combinando las posibilidades discriminativas del TDNN junto a las capacidades de modelación del locutor del HMM. La adición de un nuevo locutor se alcanza agregando un HMM, sin necesidad de reestrenar el sistema de

nuevo. Cada locutor se modela con un HMM de 6 estados entrenada con el algoritmo de Baum-Welch. Este método híbrido muestra una buena discriminación ante cada nuevo locutor, pero requiere cada vez más señal de habla para entrenar, a medida que se agregan locutores.

- *Recurrent Neural Networks* (RNN) [137]: conocidas también como redes neuronales dinámicas; se trata básicamente de una red MLP, en la que se introduce dentro de la red la naturaleza dinámica de la señal del habla, permitiéndose conexiones recurrentes de neuronas en la misma capa y entre diferentes capas. Varias aplicaciones de RNN han sido las presentadas por Artieres y otros en 1993 [138] y Hattori en 1992 [139].
- *Neural Tree Networks* (NTN) [140]: la red NTN combina las propiedades de los MLP y de los árboles de decisión. Los árboles de decisión son una colección de reglas organizadas en orden jerárquico, implementando una estructura de decisión, cada rama terminal representa una clase y cada rama no terminal representa una decisión, los nodos no terminales se implementan con perceptrones de una capa. El NTN crece recursivamente hasta que se entrene en todos los datos de entrada, los datos de prueba se aplican desde el tope de la red hasta que lleguen a una rama terminal donde se asignan a la clase de la rama. Una implementación con buenos resultados de esta ANN, es la realizada por Farell and Mammone en [141].
- *Probabilistic Neural networks* (PNN) [48]: la red PNN esta formulada como una red de cuatro capas que se entrena en un solo paso. Utiliza estimadores de función de densidad de probabilidad de Parzen que aproximan asintóticamente la densidad de los rasgos de entrada. La función de densidad de probabilidad de cada clase puede estimarse de varias formas, entre ellas, como la suma de funciones gaussianas esféricas centradas en cada vector de entrenamiento. Las PNN toman la decisión de clasificación de acuerdo con la estrategia de Bayes para las reglas de decisión. El procedimiento de entrenamiento requiere el ajuste del factor de suavizamiento, que es la desviación estándar de todas las funciones gaussianas, siendo bastante tolerante a la selección de dicho factor. El diseño no depende del entrenamiento, siendo muy rápidas de implementar, su re-entrenamiento es sencillo, solo cambiando la estadística cuando se presenten nuevos datos. Las PNN son paralelas, lo que las hace muy rápidas al clasificar. Presentan como limitantes que requieren almacenar todos los datos del entrenamiento durante la clasificación, demandando, en algunos casos, un volumen apreciable de memoria y de potencia de computo, además, al ser el factor de suavizamiento común a todas las funciones, se limita su capacidad de aprendizaje, requiriendo normalización del dato de entrada. Por último, tienen poca sensibilidad a la correlación entre entradas de datos consecutivas en el tiempo. Como solución a estas limitantes, se propone por Ganchev y otros [142] una ANN híbrida, introduciendo una RNN entre las capas de entrada y salida de la PNN, con resultados satisfactorios comparables a los obtenidos con otros clasificadores establecidos [48].

4.2.4. Modelos ocultos de Markov

Los sistemas basados en Modelos ocultos de Markov son redes de estados que intentan modelar el mecanismo de producción del habla. Están integrados por un conjunto de es-

tados (asimilables a las distintas posiciones en las que puede configurarse el tracto vocal durante una locución) que desembocan en un conjunto de posibles salidas (asimilables a las posibles distintas realizaciones alofónicas[4]).

Mientras que las aplicaciones de reconocimiento del habla presentan una topología clásica de modelado conocida como "de izquierda a derecha" (ver Fig. 6). Los modelos utilizados por los sistemas HMM para caracterizar la identidad de un locutor independiente del texto, son los denominados ergódicos[5] (ver Fig. 7), en los que no existe una ordenación correlativa[6] de las transiciones entre los distintos estados del modelo y, por lo tanto, resulta factible cualquier combinación de transición entre estados.

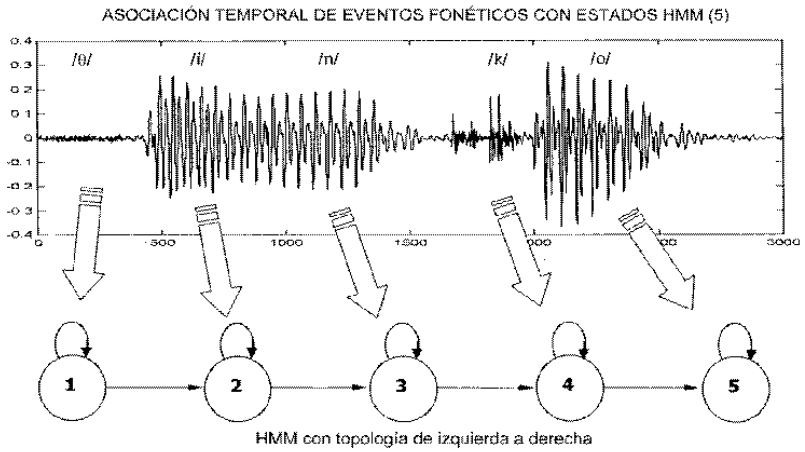


Figura 6. Modelos ocultos de Markov, topología clásica del modelo

Cada HMM viene dimensionado por tres ejes de referencia: la matriz de probabilidades de transición entre estados, el vector de probabilidades del estado inicial y la distribución de probabilidades de salida u observación. Respecto a esta última, existe la posibilidad de trabajar con HMM de observación discreta (DDHMM, por sus siglas en inglés) o de observación continua (CDHMM, por sus siglas en inglés). En general, los CDHMM modelan con mayor fidelidad y producen tasas más altas de identificación que los no continuos, si bien, en condiciones de trabajo en tiempo real los discretos parecen resultar más rápidos.

La principal ventaja de los sistemas de reconocimiento basados en HMM respecto a otros tipos ya referidos, la constituye su gran versatilidad, tanto en lo que se refiere a los procesos de entrenamiento como a ciertas características variables de la muestra: duración, contenido fonético o lingüístico, contexto, etc. A todo ello, hemos de añadir su gran adaptabilidad a la variación de las condiciones de registro o del canal de trasmisión.

[4] Se le llama alófono a cada uno de los fonos o sonidos que, en un idioma dado, se reconoce como un determinado fonema, sin que las variaciones entre ellos tengan valor diferenciativo. Cada fono corresponde a una determinada forma acústica.

[5] Hipótesis ergódica. Un sistema se denomina ergódico cuando la probabilidad de que el sistema se encuentre en un nodo determinado es constante para cualquier nodo escogido en un largo período de tiempo.

[6] Continua, seguida, sucesiva, etc.

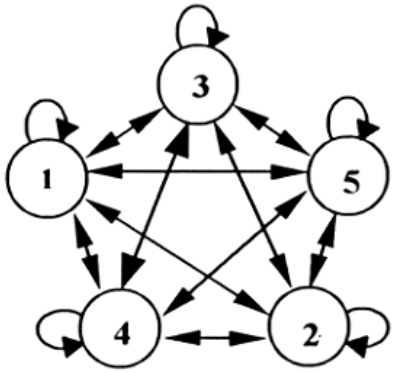


Figura 7. Modelos ocultos de Markov (ergódicos)

4.2.5. Modelos de mezclas de gaussianas

Modelan los distintos vectores de parámetros de una locución dada, realizando una suma ponderada (mezcla) de funciones de densidad de probabilidad gaussianas.

Pueden entenderse como un sistema en el que se aglutinan las virtudes de aquellos otros basados en técnicas VQ y los denominados clasificadores gaussianos uni-modales. También pudieran apreciarse como HMM de un sólo estado.

Sin embargo, utilizando GMM podemos representar, con un alto grado de fidelidad, un amplio margen de distribuciones muestrales, tales como los diferentes coeficientes cepstrales que puede generar una locución concreta.

A diferencia de los HMM ergódicos de varios estados, los GMM no precisarán, en la fase de entrenamiento, segmentar en estados ni entrenar la matriz de probabilidades de transiciones. Además, en la etapa de reconocimiento, no será necesario buscar la secuencia de estados de máxima verosimilitud (algoritmo de Viterbi), sino que bastará con acumular las probabilidades que asocia el modelo con cada uno de los vectores de entrada.

Además de las ventajas citadas, interesantes estudios comparativos sobre el rendimiento de diferentes técnicas de reconocimiento automático ante distintas circunstancias (procesos de entrenamiento, entrada de muestras, parametrización, factores de degradación, etc.), han contribuido en el uso del modelo basado en modelado GMM [8][143][144][145][146]. Esta herramienta será estudiada en los próximos epígrafes.

4.2.6. Ventajas y desventajas

A continuación se resumen las principales ventajas y desventajas de los métodos de clasificación hasta aquí expuestos.

	Ventajas	Desventajas
Dynamic Time Warping (DTW)	Detectan y comparan tramos fonéticos de alta estabilidad (vocales abiertas, consonantes nasales) aplicando técnicas de correlación cruzada, coherencia, etc, para la medida de distancias. Los sistemas DTW han sido utilizados en algunas metodologías forenses como un complemento a otros análisis clásicos [147][148].	Los principales inconvenientes de estos sistemas se relacionan con la enajenación de la información a nivel suprasegmental y la necesidad de supervisión en las tareas de segmentación.
Vector Quantization (VQ)	Reducción sensible de la capacidad de almacenamiento en el cálculo del análisis espectral y una reducción de la complejidad computacional en el cálculo de distancias (se puede usar cálculos tan simples como la distancia euclideana o la de Mahalanobis).	Sus inconvenientes más significativos están relacionados con la distorsión espectral por el error de cuantificación (al representar cada vector por un representante).
Artificial neural networks (ANN)	Las redes son robustas al ruido y permite tener en cuenta el contexto de la señal, pueden crearse redes que tengan un funcionamiento similar a las VQ, a las GMM, HMM y otros algoritmos en el reconocimiento del locutor.	Presentan como limitantes que la mayoría de las redes requieren almacenar todos los datos del entrenamiento durante la clasificación, requiriendo, en algunos casos, un volumen apreciable de memoria y poder de cálculo.
Hidden Markov Models (HMM)	Su gran versatilidad, tanto en lo que se refiere a los procesos de entrenamiento como a ciertas características variables de la muestra: duración, contenido fonético o lingüístico, contexto, etc. A todo ello, hemos de añadir su gran adaptabilidad a la variación de las condiciones de voz o del canal de transmisión y, lógicamente, su funcionalidad en condiciones dependientes de texto.	Alto costo computacional, y sus mejores resultados se encuentran en el reconocimiento del locutor dependiente del texto.
Modelos de mezclas de Gaussianas (GMM)	Las GMM pueden representar, con un alto grado de fidelidad, un amplio margen de distribuciones muestrales, como es el caso de los diferentes coeficientes cepstrales que puede generar una locución. Además de las ventajas citadas, interesantes estudios comparativos sobre el rendimiento de diferentes técnicas de reconocimiento automático, ante distintas circunstancias (procesos de entrenamiento, factores de degradación, etc.), han contribuido a tomar como mejor sistema básico a las GMM [8][143][144][145][146].	Presentan un alto costo computacional implicando un tiempo considerable al crear los modelos.

Nota

A diferencia de los HMM ergódicos de varios estados, los GMM no precisarán, en la fase de entrenamiento, segmentar en estados ni entrenar la matriz de probabilidades de transiciones. Además, en la etapa de reconocimiento, no será necesario buscar la secuencia

de estados de máxima verosimilitud, sino que bastará con acumular las probabilidades que asocia el modelo con cada uno de los vectores de entrada.

4.3. Modelos de mezclas de gaussianas para el reconocimiento del locutor independiente del texto

Han existido dos motivos principales para usar los GMM en el reconocimiento de locutores en el ámbito forense [149]:

- La idea intuitiva de que las componentes individuales de una densidad multi-modal son capaces de modelar las clases acústicas subyacentes en el proceso de verificación; esto es, que el espacio acústico que caracteriza la voz de un individuo se puede aproximar mediante un conjunto de clases acústicas (que representan conjuntos amplios de eventos acústicos), como pueden ser las vocales, las consonantes nasales o fricativas, etc. Estas clases acústicas denotan dependencia respecto a las configuraciones del tracto vocal específicas de cada locutor, siendo de gran utilidad a la hora de caracterizar a un hablante.
- La constatación empírica de los resultados alcanzados por estos sistemas en las evaluaciones internacionales llevadas a cabo por el NIST, perteneciente al Departamento de Comercio de los Estados Unidos, y en la evaluación realizada en el año 2003 por TNO-NFI (*The Netherland Forensic Institute*) con voz forense.

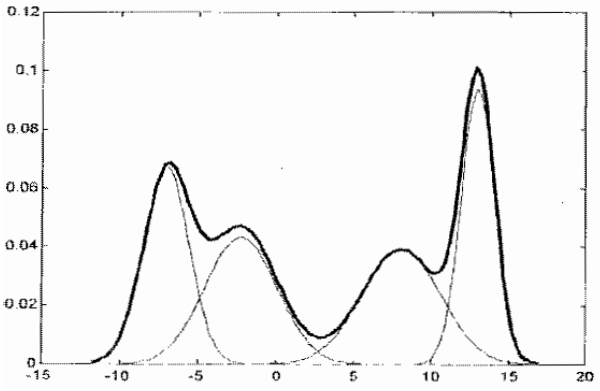


Figura 8. Curva de función de densidad de probabilidad modelada mediante una mezcla de cuatro funciones gaussianas

Los GMM modelan los distintos vectores de rasgos de una locución dada, realizando una suma ponderada (mezcla) de funciones de densidad de probabilidad gaussianas. En la Figura 8, se aprecia cómo una curva de función de densidad de probabilidad modela mediante una mezcla de cuatro funciones gaussianas una distribución experimental cualquiera. Utilizando GMM podemos representar, con un alto grado de fidelidad un amplio margen de distribuciones muestrales, tales como los diferentes coeficientes cepstrales que puede generar una locución concreta.

4.3.1. Descripción del modelo

El modelo de densidades de mezclas gaussianas es una suma pesada de M (número de mezclas) componentes de densidad descrita por:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x}), \tag{28}$$

donde:

- $\boldsymbol{x}$ es un vector D-dimensional y la matriz de rasgos

$$X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T) = \begin{pmatrix} c_{1,1} & c_{1,2} & \ldots & c_{1,T} \\ \vdots & \ddots & & \vdots \\ c_{D,1} & c_{D,2} & \ldots & c_{D,T} \end{pmatrix}$$

  es una sucesión de variables aleatorias indexadas por una variable discreta, el tiempo ($t = 1, \ldots, T$). Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y asumiremos que son independientes.
- $b_i(\boldsymbol{x})$ son las componentes de densidad, con $i = 1, 2, \ldots, M$. Cada componente es una función gaussiana de la forma:

$$b_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i),\right\} \tag{29}$$

  donde $\boldsymbol{\mu}_i$ y Σ_i son el vector de medias y la matriz de covarianza correspondientes a la i-ésima mezcla, y se obtienen de la matriz de rasgos X.
- p_i son los pesos de las mezclas con $i = 1, 2, \ldots, M$ y satisfacen la condición $\sum_{i=1}^{M} p_i = 1$.
- $\lambda = \{p_i, \boldsymbol{\mu}_i, \Sigma_i\}$ con $i = 1, \ldots, M$, representa el modelo de mezclas gaussianas donde se encuentran los parámetros que caracterizan al locutor.

Para identificar al locutor, cada uno es representado por un modelo λ de mezclas gaussianas (GMM).

Los GMM pueden tener diferentes formas, dependiendo de la elección de la matriz de covarianza:

1. El modelo puede tener una matriz por cada componente de densidad, como el modelo descrito anteriormente (covarianza por nodo).
2. Una matriz de covarianza para todas las componentes gaussianas (covarianza grande).
3. Una sola matriz para todos los modelos que representan a los locutores (covarianza global).

Características de la matriz de covarianza

La matriz de covarianza puede tener dos formas:

1. Llena.
2. Diagonal.

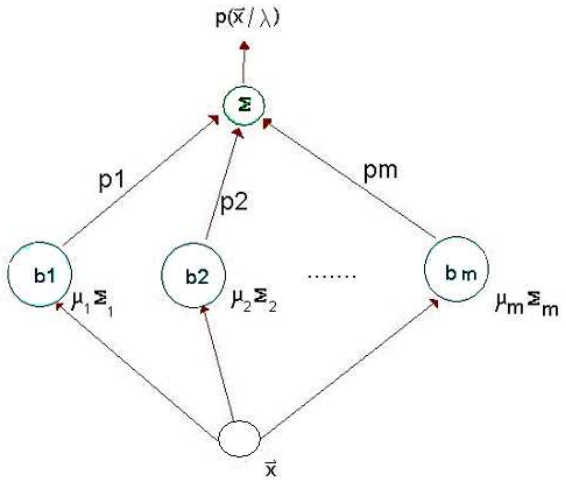


Figura 9. Descripción de un modelo de mezclas gaussianas de M componentes. Un GMM es una suma pesada de densidades, donde p_i son los pesos y b_i son las componentes gaussianas

Utilizando la diagonal, se trabaja solo con un vector que contiene la varianza. Esto se puede hacer porque la combinación lineal de la diagonal de la matriz de covarianza es capaz de modelar la correlación entre los elementos de los vectores de rasgos. El efecto de utilizar un conjunto de M matrices de covarianza llenas puede obtenerse igualmente al usar un gran conjunto de matrices diagonales de varianzas [149].

4.3.2. Estimación del parámetro de máxima verosimilitud

El objetivo del entrenamiento del modelo del locutor es estimar los parámetros de la GMM (λ). Para esto se utiliza un método llamado maximización de la verosimilitud (ML, por sus siglas en inglés), que consiste en estimar los parámetros del modelo que maximicen la probabilidad de que la GMM modele la clase, para una secuencia de T vectores de entrenamiento $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda). \tag{30}$$

Desafortunadamente, esta expresión es no lineal por el parámetro y maximizarla directamente no es posible. Sin embargo, estimar los parámetros ML puede realizarse iteradamente utilizando el caso especial de la maximización de la expectancia (EM, por sus siglas en inglés).

La idea básica del algoritmo de EM [150] es estimar, a partir de un modelo inicial λ, un nuevo modelo $\overline{\lambda}$ que cumpla con $p(X|\overline{\lambda}) \geq p(X|\lambda)$. En la próxima iteración el nuevo modelo será el inicial y el proceso se repetirá hasta que converja, $p(X|\overline{\lambda}) - p(X|\lambda) \leq \epsilon$, con $\epsilon > 0$ fijo, o cuando sobrepase la cantidad de iteraciones deseadas.

En cada iteración del EM son usadas las siguientes fórmulas de reestimación, que garantizan el crecimiento de la monotonía de la verosimilitud en el modelo:

Peso de las mezclas:

$$\overline{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda). \tag{31}$$

Medias:

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} p(i|\mathbf{x}_t, \lambda)}. \tag{32}$$

Varianzas:

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)} - \overline{\mu}_i^2. \tag{33}$$

Donde la probabilidad a posteriori para la i-ésima clase acústica (mezcla) está dada por:

$$p(i|\boldsymbol{x}_t, \lambda) = \frac{p_i b_i(\boldsymbol{x}_t)}{\sum_{k=1}^{M} p_k b_k(\boldsymbol{x}_t)}. \tag{34}$$

Dos factores críticos en el entrenamiento del GMM son la selección del orden de las mezclas y la inicialización de los parámetros del modelo *a priori* para el algoritmo de EM.

4.3.3. Identificación del locutor

Para identificar un locutor en un grupo de S locutores $\mathcal{S} = 1, \ldots, S$ representados por $\lambda_1, \lambda_2, \ldots, \lambda_S$ modelos de mezclas gaussianas se necesita encontrar el modelo del locutor que tenga la máxima probabilidad a posteriori para la secuencia del locutor dada X:

$$\hat{s} = \arg \max_{1 \leq k \leq S} \Pr(\lambda_k|X) = \arg \max_{1 \leq k \leq S} \frac{p(X|\lambda_k)\Pr(\lambda_k)}{p(X)}, \tag{35}$$

donde el miembro extremo derecho es debido a la regla de Bayes.

$p(X|\lambda_k) \rightarrow$ probabilidad de que la matriz de rasgos pertenezca al modelo λ_k, o sea, que la expresión de voz de donde se extrae la matriz de rasgos X pertenezca al locutor k.

$\Pr(\lambda_k) \rightarrow$ probabilidad del modelo k.

$p(X) \rightarrow$ la suma de las probabilidades de que la matriz de rasgos X pertenezca a todos los modelos.

Asumimos igualmente para todos los modelos que $\Pr(\lambda_k) = 1/S$ y nótese que $p(X)$ es la misma para todos también. Por tanto:

$$\hat{s} = \arg \max_{1 \leq k \leq S} p(X|\lambda_k). \tag{36}$$

Sustituyendo (30) en (36) y aplicando el logaritmo, dada la independencia entre observaciones, la verificación de los locutores queda:

$$\hat{s} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_k), \tag{37}$$

donde $p(\boldsymbol{x}_t|\lambda_k)$ fue definido en (28).

4.3.4. Detalles del algoritmo

4.3.4.1. Inicialización

Según lo indicado en la sección anterior, el entrenamiento del GMM se inicializa con un cierto modelo $\lambda^{(0)}$ y a partir de este, el algoritmo de EM garantiza encontrar un modelo de máxima verosimilitud local sin importar el punto de partida. Pero la ecuación de la probabilidad para un GMM tiene varios máximos locales [149] y diversos modelos pueden converger a diferentes máximos locales.

Para investigar el efecto de la inicialización del modelo en la verificación del locutor, se entrenaron 100 locutores usando dos métodos de inicialización:

- *Cuantificación Vectorial (VQ)*: Se clasifica la matriz de rasgos en tantas clases como mezclas gaussianas (M) sea necesarias. En este caso:

$p_i = \frac{1}{M}$	$\boldsymbol{\mu}_i$ se construye a partir de los vectores que definen los centroides de la cuantificación.	Σ_i se construye a partir de la varianza de las tramas pertenecientes a cada clase.

- *Inicialización aleatoria*:

$p_i = \frac{1}{M}$	$\boldsymbol{\mu}_i$ se construye a partir de vectores aleatorios tomados de la matriz de datos.	Σ_i se construye una matriz identidad de orden M.

Los parámetros del modelo gaussiano son:

- Matriz de rasgos 24 MFCC-Delta,
- M = 64,
- condición de parada por iteraciones 200,
- condición de parada por convergencia $p(X|\overline{\lambda}) - p(X|\lambda) \leq 0{,}001$,
- 100 locutores.

El resultado se muestra en el cuadro 5 y la Fig. 10. Se utilizaron señales de la base de datos Ahumada: dos micrófonos en diferentes sesiones M1D01 (a, b) y M7D01 (c, d, e, f) para el entrenamiento; para la prueba se utilizaron expresiones de seis sesiones distintas, dos por micrófonos y cuatro por teléfonos, (M2D01(a), M3F00(b), T3D01(c), T3F00(d), T2D01(e), T2F00(f)).

Sorprendentemente, no se encontró ninguna diferencia significativa en la verificación del locutor entre los dos métodos de la inicialización (ver Fig. 11). Los modelos obtenidos por ambos métodos pueden haber convergido a distintos máximos locales de la función de probabilidad, pero la diferencia entre sus valores de verosimilitud es insignificante.

Intuitivamente, pudiéramos preocuparnos por la cantidad de iteraciones que se requieren realizar a partir de las diferentes inicializaciones porque, como es lógico, los parámetros

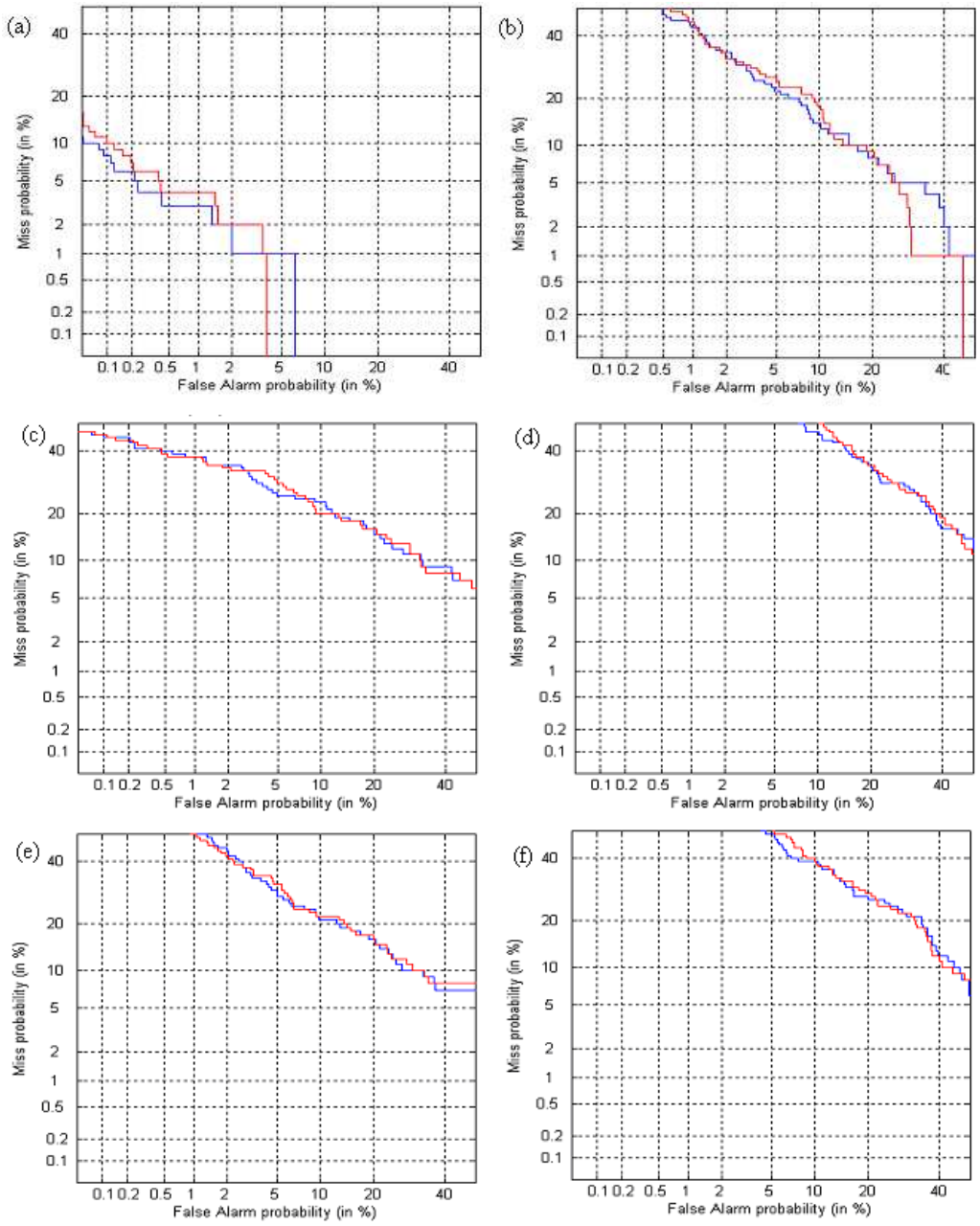


Figura 10. Dos métodos de inicialización diferentes para crear los modelos en la verificación del locutor, se entrenaron 100 locutores (Aleatoria "rojo" y Cuantificación Vectorial "azul")

Modelo	Prueba	Inicialización	
		VQ	Aleatorio
M1D01	M2D01	0.02	0.02
	M3F00	0.12	0.12
M7D01	T3D01	0.17	0.18
	T3F00	0.27	0.28
	T2D01	0.17	0.17
	T2F00	0.24	0.25

Cuadro 5. Resultados de EER en verificación para diferentes inicializaciones

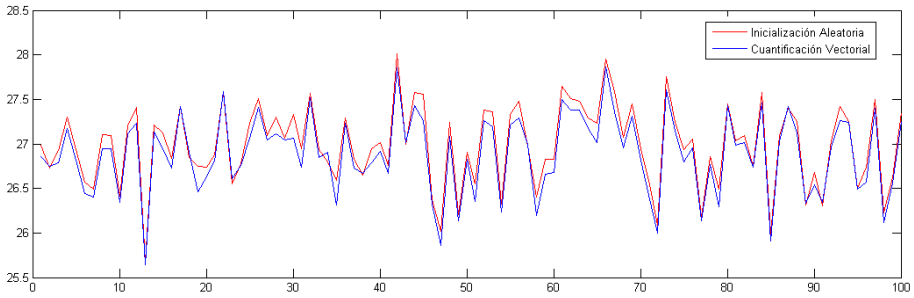


Figura 11. Comportamiento del valor de la verosimilitud de cada uno de los 100 locutores con sus respectivos modelos, para diferentes formas de inicialización

resultantes de la inicialización aleatoria se encuentran más distantes del resultado final, contrario a la inicialización, a partir del resultado de la cuantificación vectorial que deben estar ya cerca de la solución final.

Para resolver esta duda, se realizaron experimentos donde se tiene en cuenta la cantidad de iteraciones antes de converger a la solución. En estos se observó que la diferencia entre las medias es mínima: 47 iteraciones para la inicialización aleatoria y 41 para la cuantificación vectorial. Esta diferencia se equilibra al tomar en cuenta el tiempo que se utiliza en la ejecución de la cuantificación vectorial antes de comenzar las mezclas.

4.3.4.2. Orden del modelo

Determinar el número de mezclas M para el modelo del locutor es un problema importante pero difícil: "No hay manera teórica de estimar el número de los componentes de la mezcla *a priori*" [149].

Para modelar el locutor, el objetivo es elegir el número mínimo de componentes de densidades gaussianas necesarias para modelar adecuadamente un locutor, para una buena verificación.

- Elegir pocos componentes de la mezcla puede producir un modelo del locutor que no modele exactamente las características que lo distinguen en la distribución de locutores.
- Elegir demasiados componentes puede dar lugar a complejidad de cómputo excesiva en el entrenamiento y la clasificación.

En la Fig. 12 se muestra el comportamiento de la verificacion del locutor para varios órdenes del modelo (8, 16, 32, 64 y 128 componentes de densidad) y diferente duración de señal de prueba (5, 10 y 15 seg). Para el entrenamiento se tomaron muestras de 1 minuto.

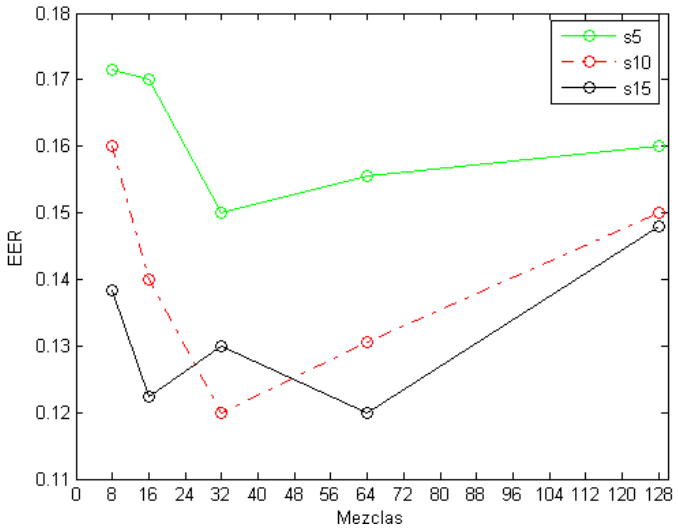


Figura 12. Error de identificacion en verificacion del locutor para diversos órdenes del modelo y tiempo de la expresión de prueba

Hay varias observaciones que se harán de estos resultados.

Primero, la clara disminución del EER de la verificación de 8 a 32 componentes de la mezcla, la estabilidad de 32 a 64 y el aumento de este de 64 a 128. La estabilidad del EER en el intervalo de 32-64 componentes de las mezclas, indica que este es el número de componentes necesarios para modelar locutores para la base de datos utilizada y con un tiempo aproximado de un minuto de expresión de duración de las muestras. O sea los modelos deben contener, por lo menos, este número mínimo de componentes y esta duración para mantener un buen funcionamiento en la verificación del locutor.

En el experimento con expresiones de pruebas crecientes respecto al tiempo (5s, 10s 15s), el EER disminuyó al incrementar el tiempo para todos los modelos con 8, 16, 32, 64 y 128 componentes de densidad. Esto sugiere que por lo menos 15s o más de voz es necesario para lograr una buena verificación del locutor.

Este reporte ha evaluado el uso de los GMM para la verificación robusta del locutor independiente del texto. Los GMM fue evaluado específicamente para la verificación usando expresiones de voz que variaron en: voces leídas, espontáneas, variabilidad del canal, variabilidad en la duración de las expresiones y cambio de sesión.

La evaluación experimental realizada examinó varios aspectos del uso de los GMM para la verificación del locutor independiente del texto. Algunas observaciones:

- Los GMM son insensibles al método de inicialización que se utilice para crear el modelo, logrando la misma verificación del locutor.

Algoritmo 1: Algoritmo para el entrenamiento de los GMM

Input: $\lambda = \{p_i, \boldsymbol{\mu}_i, \Sigma_i\}$, con $i = 1, \ldots, M$
n: cantidad de iteraciones;
ϵ: error entre una iteración y la próxima
Output: $\overline{\lambda}$ tal que $p(X, \overline{\lambda}) \geq p(X, \lambda)$
while *iteraciones* $\leq n$ **do**
  estimar los parámetros del modelo $\overline{\lambda} = \{\overline{p}_i, \overline{\boldsymbol{\mu}}_i, \overline{\sigma}_i^2\}$
  if $p(X, \overline{\lambda}) - p(X, \lambda) \geq \epsilon$ **then**
    $\lambda = \overline{\lambda}$
  else
    break
  end
end

- La limitación de la varianza es importante en el entrenamiento para evitar las singularidades del modelo.
- El orden de mezclas mínimo para crear el modelo oscila de 32 a 64 componentes para modelar adecuadamente los locutores y para alcanzar un buen resultado en la verificación.
- Los GMM mantienen un alto desempeño en la verificación con el aumento del tamaño de la población.
- Los GMM tienen la capacidad de modelar datos con densidades arbitrarias.
- Con los GMM se obtienen mejores resultados experimentales que con técnicas usadas con anterioridad como la cuantificación vectorial, redes neuronales y modelos de Markov.

Estos resultados indican que los GMM proporcionan una representación robusta del locutor para la difícil tarea de la verificación a partir de locuciones independientes del texto. Los GMM se pueden poner en ejecución, fácilmente, en una plataforma en tiempo real [149][152].

4.3.5. Modelos de background

Dado un segmento de habla X y un locutor hipótesis S, la tarea del verificación del locutor consiste en determinar si X fue hablada por S. Esta se puede ver como la decisión entre dos hipótesis:

- H_0: X es del locutor S.
- H_1: X no es del locutor S (hipótesis alternativa).

La prueba óptima a decidir entre estas dos hipótesis es una probabilidad dada por el cociente [7]:

$$\frac{p(X|H_0)}{p(X|H_1)} \begin{cases} > \Theta, \text{ acepta } H_0, \\ < \Theta, \text{ acepta } H_1. \end{cases} \tag{38}$$

donde $p(X|H_0)$ es la función de densidad probabilística (verosimilitud) para evaluar la hipótesis H_0 en la observación de voz X y la función de verosimilitud para la hipótesis H_1 es $p(X|H_1)$. El umbral de decisión para aceptar o rechazar H_0 es Θ.

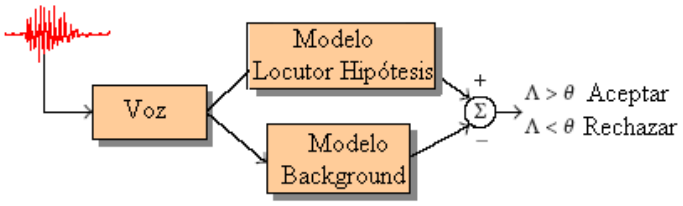


Figura 13. Componentes básicos para la detección del locutor a partir de la Ecuación (38)

La matriz de rasgos de X se utiliza entonces para calcular la verosimilitud de H_0 y H_1. Se denota el modelo que representa a H_0 por λ_{hyp}, que caracteriza al presunto locutor S.

Por ejemplo, se puede asumir que una distribución gaussiana representa lo mejor posible la distribución de los vectores de rasgos para H_0, de modo que λ_{hyp} contenga los parámetros, vector de pesos, vector de medias y la matriz de covarianza de dicha distribución y el modelo $\overline{\lambda}_{hyp}$ representa la hipótesis alternativa H_1. Entonces, la verosimilitud está dada por $p(X|\lambda_{hyp})/p(X|\overline{\lambda}_{hyp})$. Para eliminar la división utilizamos el logaritmo, quedando como:

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\overline{\lambda}_{hyp}). \tag{39}$$

Mientras que el modelo para H_0 está bien definido y se puede estimar usando la expresión de voz del entrenamiento de S, el modelo para $\overline{\lambda}_{hyp}$ está menos bien definido, puesto que, potencialmente, debe representar el espacio entero de alternativas posibles al locutor supuesto.

Dos acercamientos principales se han tomado para modelar esta hipótesis. El primer acercamiento es utilizar un conjunto de modelos de locutores donde *no se encuentre el locutor a reconocer*, para cubrir el espacio de la hipótesis alternativa. En varios contextos, este sistema de otros modelos de locutores se ha llamado verosimilitud de conjunto [40], cohortes [40][23], y modelos de background [40][116]. Tomando un conjunto de N modelos de locutores de background, el modelo alternativo de la hipótesis es representado por:

$$p(X|\overline{\lambda}_{hyp}) = f(p(X|\lambda_1), \ldots, p(X|\lambda_N)), \tag{40}$$

donde $f(\cdot)$ es el promedio o máximo de los valores de probabilidad del conjunto de locutores de background. La selección, el tamaño y la combinación de los locutores de background han sido el tema de muchas investigaciones [40][23][116][29]. Generalmente, obtener el mejor funcionamiento con este acercamiento, requiere el uso de conjuntos de locutores *específicos* para el modelo de background. Esto puede ser una desventaja usando una gran cantidad de presuntos locutores, pues cada uno requiere su propio conjunto de modelos de background.

El segundo mejor acercamiento al modelo de la hipótesis alternativa es reunir una sola expresión de voz de varios locutores y entrenar un solo modelo. Varios términos utilizados para este modelo son: modelo general [117], modelo del mundo, y modelo universal de background (UBM, por sus siglas en inglés) [10]. Dada una colección de muestras de expresiones de voz de una gran cantidad de representantes (locutores) de la población esperada durante la verificación, un solo modelo λ_{bkg} (UBM) se entrena para representar la hipótesis alternativa. Las investigaciones sobre este acercamiento se han centrado en la selección de

los locutores y de la expresión de voz a utilizar en el entrenamiento del modelo universal de background [10][118]. La ventaja principal de este acercamiento es que el entrenamiento se realiza una sola vez, obteniendo un UBM que será utilizado para todos los locutores presumidos en el reconocimiento. El uso de un solo modelo de background se ha convertido en el acercamiento más usado en sistemas de verificación del locutor.

En un experimento realizado con UBM para normalizar la verosimilitud, se utilizaron 500 expresiones de voz de 50 locutores (30 minutos de voz) y se modeló una GMM de 256 mezclas para crear el UBM. Los resultados de la clasificación se muestran en la figura 14.

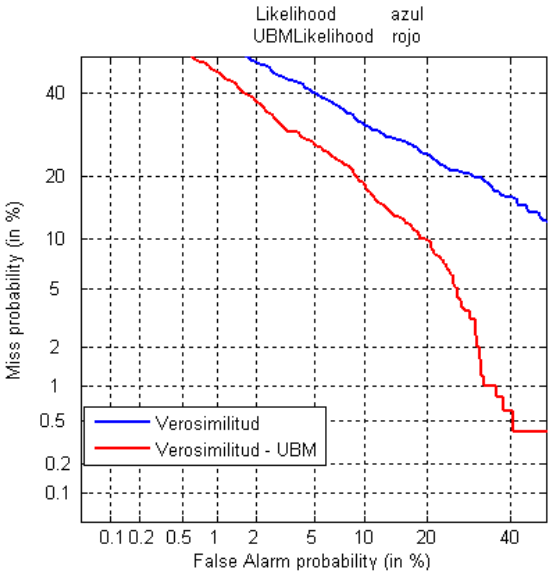


Figura 14. Resultados de la verificación de locutores con expresiones de voz sobre líneas telefónicas sin/con normalización de la verosimilitud UBM

El uso de los modelos de background como una normalización a la verosimilitud obtenida a partir de las GMM mejoran sustancialmente los resultados, como se muestra en la figura anterior.

4.3.6. Modelos de mezclas de gaussianas adaptadas

La idea básica de las GMM-adaptadas consiste en obtener el modelo del locutor mediante la adaptación de los parámetros[7] del UBM utilizando los rasgos del locutor. A este entrenamiento se le llama adaptación Bayesiana o estimación por máximo a posteriori (MAP, por sus siglas en inglés) [150]. Diferente al acercamiento estándar de maximizar la verosimilitud de un modelo para el locutor que sea independiente del UBM, este nuevo enfoque llamado adaptación consiste en obtener el modelo del locutor mediante la actualización de los parámetros bien entrenados en el UBM. Esto proporciona una conexión

[7] Cuando se habla de los parámetros de UBM se refiere a los pesos, las medias y las varianzas del modelo.

más estrecha entre el modelo del locutor y el UBM, que no solo provoca un mejor funcionamiento que los modelos independientes, sino que también tiene en cuenta una técnica rápidapara calcular la verosimilitud. Como el algoritmo de EM, la adaptación es un proceso de estimaciónde dos etapas. El primer paso es idéntico al primer paso del algoritmo de EM, donde se calculan las estimaciones de las estadísticas suficientes[8] de los datos de entrenamiento del locutor para cada mezcla en el UBM. La diferencia se encuentra en el segundo paso del algoritmo de EM. Para la adaptación, las "nuevas" estimaciones de los parámetros del modelo del locutor se combinan con las estimaciones de los parámetros del modelo de background "viejo" usando un coeficiente para la combinación que es dependiente de los datos. El coeficiente que se utiliza para la combinación de los parámetros tiene que garantizar que las mezclas con altas verosimilitudes con los datos del locutor influyan más que las mezclas del modelo de background, en la actualización de los parámetros y en las mezclas con bajas verosimilitudes con los datos del locutor influyan más las mezclas del modelo de background.

La adaptación consiste en tomar un UBM y los rasgos del locutor que clama su identidad, $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, luego determinamos la alineación probabilística de los rasgos de entrenamiento en los componentes o mezclas del UBM. Es decir, para la i-ésima mezcla en el UBM, calculamos

$$\Pr(i|\boldsymbol{x}_t, \lambda) = \frac{p_i b_i(\boldsymbol{x}_t)}{\sum_{k=1}^{M} p_k b_k(\boldsymbol{x}_t)}. \tag{41}$$

Entonces utilizamos $Pr(i|\boldsymbol{x}_t, \lambda)$ y $\boldsymbol{x}_t$ para calcular los parámetros del modelo (el peso, la media, y la varianza):

- Peso de las mezclas:

$$n_i = \sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda). \tag{42}$$

- Medias:

$$E_i(\boldsymbol{x}) = \frac{1}{n_i} \sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t. \tag{43}$$

- Varianzas:

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)x_t^2. \tag{44}$$

Hasta aquí el primer paso, que es el mismo que la expectancia en el algoritmo de EM. Finalmente, los nuevos parámetros se utilizan para actualizar los parámetros del UBM de la i-ésima mezcla para crear los parámetros adaptados de ésta, con las siguientes ecuaciones:

- Peso de las mezclas adaptadas:

$$\hat{w}_i = [\alpha_i^w n_i/T + (1 - \alpha_i^w)w_i]\gamma. \tag{45}$$

- Medias adaptadas:

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\boldsymbol{x}) + (1 - \alpha_i^m)\boldsymbol{\mu}_i. \tag{46}$$

[8] Éstas son las estadísticas básicas requerida para calcular los parámetros deseados. Para un GMM, el peso, la media y la varianza.

- Varianzas adaptadas:

$$\sigma_i^2 = \alpha_i^v E_i(x^2) + (1 - \alpha_i^v)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2. \tag{47}$$

El factor de escala γ es utilizado en los pesos de las mezclas adaptadas para asegurar que las sumas de ellos sea la unidad. Los coeficientes de la adaptación que controla el balance entre las viejas y nuevas estimaciones son $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ para los pesos, las medias y las varianzas, respectivamente.

Para cada parámetro de cada mezcla, el coeficiente dependiente de los rasgos α_i^ρ, $\rho \in \{w, m, v\}$, se utiliza en las ecuaciones anteriores (45,46,47) y se define como:

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho}, \tag{48}$$

donde r^ρ es un factor fijo de "relevancia".

Usar un coeficiente de adaptación dependiente de los rasgos permite una adaptación de los parámetros dependiente de las mezclas. Si un componente de las mezclas tiene una probabilidad n_i baja en los nuevos datos, entonces $\alpha_i^\rho \to 0$, causando poco énfasis de los nuevos parámetros –potencialmente no entrenado– y enfatizando en los viejos parámetros –mejor entrenados. Para los componentes de las mezclas con altas probabilidades $\alpha_i^\rho \to 1$, causando mayor énfasis en los nuevos parámetros y poco en los viejos. El factor de relevancia controla cuántos nuevos datos deben ser observados en una mezcla antes de que los nuevos parámetros reemplacen los viejos parámetros. Este enfoque debe ser robusto, con datos de entrenamiento limitados.

El uso de factores de relevancia parámetro-dependientes permiten una mejor sintonía para diferentes adaptaciones en cuanto a los pesos, las medias, y las varianzas. Sin embargo, experimentos divulgados en [153] demostraron que el uso de varios coeficientes para la adaptación tiene poco beneficio respecto al uso de un solo coeficiente. Reynolds [154] utiliza, en un sistema de GMM-UBM un solo coeficiente de adaptación para todos los parámetros

$$\alpha_i^w = \alpha_i^m = \alpha_i^v = n_i/(n_i + r), \tag{49}$$

con un factor de relevancia $r = 16$.

Puesto que la adaptación es dependiente de los rasgos, no todas las gaussianas en el UBM son adaptadas durante el entrenamiento del modelo del locutor. Conocer la cantidad de gaussianas no adaptadas puede ser un factor importante en los requisitos para el almacenaje de los modelos de locutores, reduciendo el espacio en memoria, puesto que es posible almacenar modelos usando solamente la diferencia con el UBM.

Resultados publicados en [154][155] usando los datos NIST SRE (*NIST Speaker Recognition Evaluations*) [156] y la comparación de otros sistemas que no usan la adaptación muestran que el enfoque de la adaptación proporciona resultados superiores a los resultados obtenidos en los sistemas independientes. Una posible explicación para estos resultados es que la verosimilitud obtenida a partir de los modelos adaptados no esta afectada por acontecimientos acústicos no vistos en el reconocimiento de la voz.

Los *índices de verosimilitud* para una secuencia de prueba de los vectores de característica X se calculan como:

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{UBM}). \tag{50}$$

El hecho de que el modelo del supuesto locutor fuera adaptado del UBM, permite un método más rápido para calcular la verosimilitud que la evaluación de forma estándar de los GMM, como en la Ecuación (37). Esto sucede producto de dos efectos observados. El primero es que cuando un GMM suficientemente grande se evalúa para una matriz de rasgos, sólo algunas de las mezclas contribuyen significativamente al valor de la verosimilitud. Esto es porque el GMM representa una distribución sobre un espacio grande, mas es un solo vector que estará cerca solamente de algunos componentes del GMM. Así, los valores de la verosimilitud se pueden aproximar muy bien usando solamente los C mejores componentes de las mezclas. El segundo efecto observado es que los componentes del GMM adaptado conservan una correspondencia con las mezclas del UBM, de modo que los vectores cerca de una mezcla particular en el UBM también estén cerca de la mezcla correspondiente en el modelo del locutor.

Al usar estos dos efectos para ganar velocidad de cálculo el algoritmo queda de la siguiente manera: Para cada vector de rasgos, determinar las mezclas (C) que mejoran la verosimilitud en el UBM y calcular la verosimilitud de UBM usando solamente éstas C mezclas. Después, se utiliza el vector de rasgos solamente contra las C componentes correspondientes al modelo del locutor adaptado, para evaluar la probabilidad del locutor. Para un UBM con M mezclas, este enfoque requiere solamente $M + C$ cálculos gaussianos por cada vector de rasgos, comparado con los $2M$ cálculos gaussianos para la probabilidad normal. Cuando hay múltiples modelos de locutores supuestos para cada segmento de la prueba, los ahorros llegan a ser mucho mayor. En el sistema de GMM-UBM descrito en [155], Reynolds utiliza un valor de $C = 5$.

En el año 2004, F. Bimbot, J.F. Bonastre, D. A. Reynolds y otros plantean que [157]: "la tecnología del reconocimiento automático del locutor parece haber alcanzado un nivel suficiente de madurez para la aplicación realista en el ámbito de la ciencia forense."

Para ese entonces se trabajaba en sistemas automatizados para el reconocimiento del locutor que utilizan como rasgos los MFCC con normalización, ya sea CMN o CMVN, y otros como los RASTA [65], modelación con GMM, adaptación MAP, normalización del score con el UBM. Esta es la línea base de los algoritmos en los sistemas de reconocimiento automático del locutor en este momento.

4.4. Normalización de los resultados de la comparación

4.4.1. Objetivos de la normalización del resultado

El último paso en la verificación del locutor es la toma de decisión. Este proceso consiste en puntuar la probabilidad resultante de la comparación entre el modelo del locutor reclamante y la señal de habla de entrada, y comparar con un umbral de decisión. Si la puntuación es más alta que un umbral, el locutor reclamante será aceptado, si no, rechazado.

La obtención de los umbrales de decisión es un problema abierto en la verificación del locutor, siendo fijado generalmente de forma empírica y su confiabilidad no puede ser asegurada mientras el sistema esté funcionando.

Tanto en la obtención de la puntuación de la semejanza como en la obtención del umbral existe determinada variabilidad, un hecho bien conocido, la cual proviene de diversas fuentes:

- La naturaleza del habla para el entrenamiento puede variar entre locutores, las diferencias pueden provenir del contenido fonético, la duración, el ruido del ambiente, así como la calidad del entrenamiento de los modelos. Dicha variabilidad inter-locutor, no intrínseca de cada locutor, tiene que ser considerada como un factor potencial que afecta la confiabilidad de los límites de la decisión.
- Las desigualdades entre el habla del entrenamiento para crear el modelo del locutor y el habla de prueba del mismo locutor; dos factores pueden contribuir: la variabilidad intra-locutor (variación en la voz del locutor debido a la emoción, estado de la salud y edad) y los cambios de las condiciones del ambiente en el canal de transmisión, el material de grabación o el ambiente acústico.
- La naturaleza y calidad de los segmentos de habla para el entrenamiento y la prueba influyen en el valor de los resultados.

4.4.2. Técnicas de normalización

La normalización de la puntuación en los resultados se ha introducido explícitamente para hacer frente a la variabilidad de los resultados y para sintonizar más fácil el umbral de decisión. Las técnicas de normalización, básicamente, consisten en centrar la distribución de la puntuación generada por el sistema. Lo que se traduce en obtener una estadística de normalización (media y desviación estándar) y en restarle la media a la puntuación de la comparación entre el modelo del locutor reclamante y la señal de habla de entrada y luego dividirlo entre la desviación estándar. Aunque existen variadas técnicas de normalización de la puntuación, mencionaremos las más extendidas en su uso [157]:

- **Z-norm**: la técnica *Zero-normalization* se ha utilizado masivamente en la verificación del locutor desde los años noventa. En la práctica, cada modelo del locutor se compara contra un grupo de señales de voz producidas por algún impostor, dando por resultado una distribución de las puntuaciones de la semejanza del impostor. La estadística de dicha distribución de puntuaciones del impostor - media y desviación estándar- normaliza las puntuaciones de la semejanza del sistema de verificación del locutor. Z-norm se considera una técnica de normalización dependiente del locutor. Una de las ventajas de este método es que la obtención de la estadística para la normalización se realiza fuera de línea durante el entrenamiento de los modelos de locutores.
- **H-norm**: para voz por teléfono, la mayoría de los modelos de locutores responden de forma diferente, según el tipo del microteléfono usado durante la realización de la prueba. Se propone una variante de la técnica Z-norm, denominada *Handset-normalization*, para enfrentar las diferencias del microteléfono entre el entrenamiento y la prueba. Aquí, la estadística para la normalización se estima comparando cada modelo del locutor contra las señales de voz producidas por impostores pero dependiente de cada tipo de microteléfono usado. Durante la prueba, el tipo de microteléfono referente a la señal de voz a probar, determina la estadística a utilizar para la normalización de la puntuación. Este método posee la misma ventaja que el anterior.

- **T-norm**: la técnica *Test-normalization* difiere de Z-norm por el uso de modelos del impostor en vez de señales de voz, en la prueba. Durante la prueba, la señal de voz no sólo se compara con el modelo del locutor reclamante, sino con un grupo de modelos de impostores previamente entrenados, para estimar consecutivamente la puntuación de la distribución de impostores y la estadística de la normalización para la puntuación. Si Z-norm se considera una técnica de normalización dependiente del locutor, T-norm es dependiente de la prueba. Como la misma expresión de prueba se utiliza para la prueba contra el modelo del locutor y la estimación de la estadística de normalización, T-norm evita el efecto de posibles diferencias entre las expresiones de prueba y las expresiones para obtener la estadística de la normalización, lo cual puede ocurrir con Z-norm. T-norm tiene que ser realizado en línea durante la prueba.

- **HT-norm**: basado en la misma observación que H-norm, se ha propuesto una variante de T-norm, nombrado *Handset Test-normalization*, que tiene en cuenta la información del microteléfono durante la prueba. Aquí, la estadística de normalización dependiente del microteléfono se estima probando cada señal de voz contra modelos de impostores, dependientes del microteléfono. Durante la prueba, el tipo de microteléfono referente al modelo reclamado del locutor determina la estadística a utilizar para la normalización de la puntuación.

- **C-norm** : fue introducido en la verificación del locutor sobre celulares, cuando la voz viaja por teléfonos portátiles con microteléfonos no identificados. Para hacerle frente, se propone un agrupamiento a ciegas de las voces para la normalización, seguida de un proceso similar a H-norm, considerando cada grupo como un microteléfono distinto.

4.5. Evolución en términos de mejora del reconocimiento automático del locutor desde el 2004 hasta hoy

En este epígrafe se ilustra brevemente la evolución en el estado del arte del RAL. Para esto, se utilizan como línea base los algoritmos que se basan en modelos de mezclas gaussianas GMM-UBM adaptados, que alcanzaron el estado del arte durante la NIST'04 SRE.

Como indica el crecimiento del número de participantes en las evaluaciones internacionales NIST [158], un creciente interés ha surgido en el reconocimiento del locutor en los últimos años. Al mismo tiempo, las mejoras significativas en el rendimiento han logrado más o menos reducir a la mitad la tasa de error en las tres últimas campañas NIST.

Respecto a la evolución de los algoritmos hasta la actualidad, en términos de mejorar la robustez ante la variabilidad y precisión, se pueden mencionar tres principales aportes: las máquinas de soporte vectorial (SVM) [159] [160], proyección de los atributos con ruido (NAP) [77] y factor de análisis (FA) [161][73], de los cuales se puntualizarán solo las SVM en este reporte.

4.5.1. Máquinas de soporte vectorial para el reconocimiento del locutor independientes del texto

Las máquinas de vectores de soporte (*Support Vector Machines*, SVM) han sido usadas en diversos campos incluyendo reconocimiento de imágenes, categorización de textos e investigaciones de ADN, con muy buenos resultados. Hoy en día, esta potente herramienta ha sido introducida también en el área de reconocimiento del locutor, presentando resultados sobresalientes al fusionarlas o combinarlas con las GMM.

Las SVM ofrecen una estrategia alternativa en la clasificación del locutor para los ampliamente usados GMM y ha sido investigada por muchos para la verificación del locutor [159,162,163,164], también la SVM puede trabajar en espacios no lineales, utilizando un kernel que hace un mapeo de los patrones de entrada X en algún espacio de características donde se comporten de forma lineal.

Las SVM constituyen un clasificador discriminativo que modela las fronteras entre las clases, a diferencia de las GMM que son un clasificador generativo que modela las distribuciones de probabilidad de las clases.

Puesto que la SVM es un clasificador de dos clases, Campbell [159] maneja el reconocimiento del locutor como una verificación. Si tenemos un conjunto cerrado de locutores y se quiere realizar una identificación, entonces se entrena un modelo del locutor objetivo y un modelo de background con los demás locutores.

4.5.1.1. Kernel secuenciales y SVM lineales

Los kernels secuenciales están definidos como:

$$K(X,Y) = \phi(X)^T R^{-1} \phi(Y), \tag{51}$$

donde $\phi(X)$ es un vector de rasgos de alta dimensión que representa la expresión X (una observación o trama), $\phi(Y)$ vector de rasgos de alta dimensión de la expresión Y y R^{-1} una matriz de normalización diagonal.

Siguiendo a grandes rasgos el entrenamiento y las pruebas descritas en [159], una SVM lineal es suficiente, con vectores de entrada $R^{-1/2}\phi(X)$.

Recordando que la verificación es un problema binario, entonces todos los vectores de rasgos correspondientes a un determinado locutor en el entrenamiento son etiquetados, por ejemplo, con 1 y son comparados con vectores de rasgos del impostor, con la etiqueta -1, obteniéndose como resultado del entrenamiento una definición del hiperplano que separa ambas clases.

$$f(x) = \sum_{i=1}^{N_{sv}} \alpha_i t_i R^{-1/2} \phi(X_i) x + d, \tag{52}$$

donde N_{sv} es la cantidad de vectores de soportes, t_i es la salida ideal, $\sum_{i=1}^{N_{sv}} \alpha_i t_i = 0$ y d es el desplazamiento del hiperplano al centro de coordenadas. Este clasificador puede ser compactado para que permita la evaluación de $f(x)$ como un simple producto escalar:

$$w_x = \begin{bmatrix} \sum_{i=1}^{N_{sv}} \alpha_i t_i R^{-1} \phi(X_i) \\ d \end{bmatrix}, \tag{53}$$

obteniéndose como función de decisión:

$$Score_{SVM}(X,Y) = f(R^{-1/2\phi(Y)}) = \phi(Y)^T w_x. \tag{54}$$

La descripción anterior puede ser utilizada en diferentes enfoques de la SVM en verificación, lo cual sólo requeriría la definición del kernel en (51), a través de:

$$\begin{cases} X \rightarrow \phi(X), \\ r^{-1/2}. \end{cases}, \tag{55}$$

donde r es la diagonal de matriz R con dimensión E, y E se refiere al tamaño de la expansión de la función ϕ.

4.5.1.2. Kernels secuencial con discriminante lineal generalizado (GLDS)

Para el GLDS (de sus siglas en inglés), como se describe en [159] la expansión polinomial $\phi_p(f_t)$ está calculada.

Esto conlleva a un vector de dimensión $\binom{F+p}{p}$ formado por todos los posibles monomios[9] de grado p a partir del vector de rasgos original f_t de tamaño F.

El promedio de la expansión polinomial sobre las T_X tramas en la expresión X esta dado por

$$\phi_{GLDS}(X) = \overline{\phi}_p(X) = \frac{1}{T_X} \sum_{t=1}^{T_X} \phi_p(f_t), \tag{56}$$

y es calculado en la expresión del impostor con N_c coeficientes y T_n tramas.

$$r_{GLDS} = \frac{1}{N_c} \sum_{n=1}^{N_c} \frac{1}{T_n} \sum_{i=1}^{T_n} \phi_p(f_i)\phi(f_i). \tag{57}$$

Cada componente de r es la media de los valores cuadrados sobre todas las expansiones del impostor. Una observación interesante es que r captura la dinámica de los coeficientes del polinomio.

4.5.1.3. Kernel linear con GMM supervector (GSL)

El desarrollo de una métrica en el espacio de las GMM [165] llevó a la idea de utilizar nuevos kernels secuenciales basados en GMM supervectores [164].

Una nueva expansión se define por:

$$\phi_{GSL}(X) = m_X = \begin{bmatrix} m_X^1 \\ \vdots \\ m_X^i \\ \vdots \\ m_X^G \end{bmatrix}, \tag{58}$$

[9] Un monomio es un tipo particular de polinomio, con un único término. De hecho, todo polinomio es una suma de monomios.

el cual es el supervector de valores de medias m_X^i tomadas del modelo obtenido por la GMM entrenadas en la expresión X. Cada GMM tiene G componentes gaussianas, cada una con un vector de rasgos acústicos de C coeficientes, esto da un $\phi_{GSL}(X)$ de tamaño GC.

Los parámetros de peso (λ_i) y varianza del UBM son usados para definir r como:

$$r_{GSL}^{-1/2} = \begin{bmatrix} \sqrt{\lambda_1}\Sigma_1^{-1/2} \\ \vdots \\ \sqrt{\lambda_i}\Sigma_i^{-1/2} \\ \vdots \\ \sqrt{\lambda_G}\Sigma_G^{-1/2} \end{bmatrix}. \tag{59}$$

Es posible considerar otros kernels secuenciales en este marco. En [166], son probados con éxito kernels basados en los pesos (Fisher Kernel).

En [169], se compara el comportamiento de la clasificación con GMM, y con SVM con kernels GLDS y GSL. Se utiliza la misma configuración de rasgos y la base NIST'04. Se obtienen modelos con $G = 512$ componentes gaussianas y $C = 50$ coeficientes acústicos, obteniéndose un supervector de medias de dimensión 25600.

Para el sistema GLDS, una expansión de grado $p = 3$ fue usada, obteniendo un supervector de tamaño $\binom{50+3}{3} = 23426$.

La figura 15 muestra que el SVM-GSL y el supervector de medias obtenido de las GMM adaptadas dan los mejores resultados.

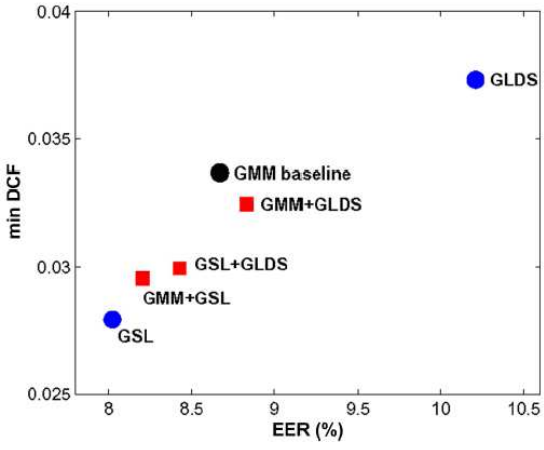


Figura 15. Muestra el funcionamiento en términos de igualdad de la tasa de error (EER) y el valor mínimo de la función de detección de costo (DCF), de los algoritmos GMM, GSL, GLDS y la fusión entre ellos

Métodos de clasificación. Conclusiones

Los métodos de clasificación han evolucionado en el tiempo. En la décadas del 60 y 70 predominó la clasificación comparando plantillas de palabras, en los 80 se clasificó aplicando la distorsión dinámica en el tiempo (DTW) y la cuantización vectorial (VQ). A partir de los finales de los 1990 y principios del 2000 se desarrollan enfoques estadísticos de clasificación, como los modelos ocultos de Markov (HMM), las mezclas gausianas (GMM), y los modelos universales de background (UBM), lográndose mejores resultados de clasificación.

A principios del 2000, comienzan a aplicarse diferentes formas de adaptar los modelos GMM de los locutores a partir de los modelos universales UBM, predominando la adaptación máxima *a posteriori* (MAP), que ha sido clave en las mejoras de los clasificadores en los últimos años, alcanzando el estado del arte de dichos modelos durante la evaluación NIST'04.

A partir de aquí se comienza a desarrollar un enfoque de clasificación discriminativa partiendo de los modelos GMM, como los GMM *Supervector Linear Kernel* (GSL) que utilizan supervectores construidos a partir de las medias de los modelos GMM adaptados MAP para clasificar los locutores con máquinas de soportes vectoriales (SVM).

A continuación, un breve resumen de lo sucedido en el estado del arte o funcionamiento de los algoritmos en las últimas campañas NIST. En cada caso, se considera sólo un único sistema, evitando así la necesidad de examinar los posibles beneficios de la fusión. Una visión general de los resultados de la evaluación del reconocimiento del locutor (SRE) está disponible en [168]. El concepto de estado del arte es algo subjetivo, pero en este, caso si se toma como definición los mejores resultados producidos, para esto utilizaremos los valores del EER y minDCF en la decisión del mejor funcionamiento, podemos ilustrar la evolución de la siguiente manera:

- **04 SRE**: GMM estándar con T-norm [169] (EER = 9.92 %; minDCF = 0.0425);
- **05 SRE**: GMM adaptados y fusionados con SVM. (EER = 7.20 %; minDCF = 0.0335);
- **06 SRE**: NAP [77] y FA [161][73] con GSL-NAP. (EER = 5.02 %; minDCF = 0.0226);

En la Fig. 16 se observan los resultados obtenidos en [74].

Todos estos algoritmos y las modificaciones introducidas han provocado en los últimos años grandes pasos de avance basados en los índices del EER, que es el criterio utilizado para evaluar tanto el funcionamiento de los algoritmos como el progreso en los sistemas de reconocimiento del locutor, (línea base en el 2004, EER = 9.92 %; sistemas fusionados en el 2005, EER = 7.20 %; NAP, FA y GSL-NAP en el 2006, EER = 5.02 %; GSL-FA-unsupervised en el 2007, EER = 2.27 %). A pesar de esto, el problema de la clasificación sigue siendo un desafío clave para el reconocimiento del locutor.

5. Conclusiones

El presente reporte de investigación recoge los resultados del estudio del estado del arte realizado por los autores en el marco del proyecto fundamental "Métodos robustos de procesamiento del habla", que se lleva a cabo en el CENATAV.

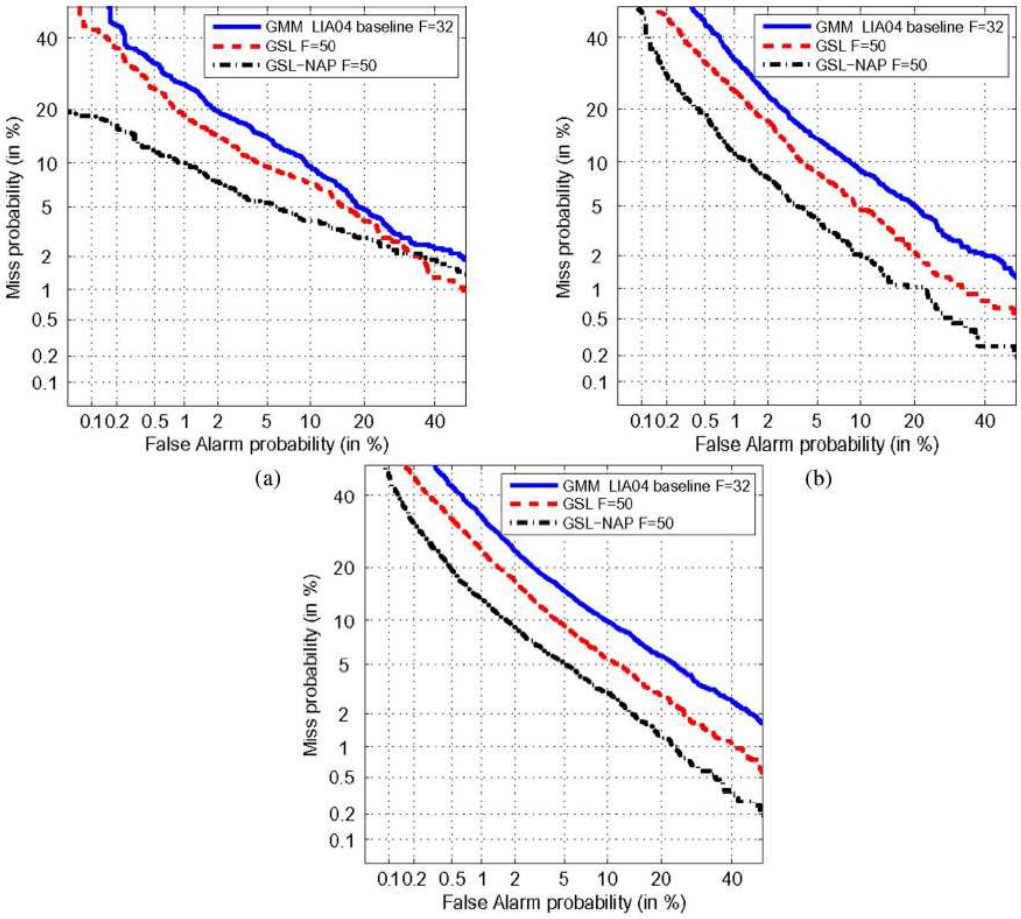


Figura 16. Se utilizaron tres bases de datos: (a) NIST'05, (b) NIST'06 locutores masculinos solamente, (c) NIST'06 locutores masculinos y femeninos. Se compararon las GMM, GSL, y GSL-NAP

El Reporte hace énfasis en el análisis de los métodos de extracción, selección y clasificación de rasgos acústicos mas comúnmente usados en el reconocimiento del locutor y del habla en general, pudiendo servir como referencia para cualquier especialista del tema. Posee además una amplia y actualizada bibliografía sobre procesamiento de habla.

El estudio del estado del arte conllevó a determinar áreas de investigación donde se presentan aún problemas a resolver:

- La extracción de rasgos acústicos con una mayor robustez ante la variabilidad del canal y del locutor, y ante el ruido,
- la selección de rasgos acústicos mas representativos y menos redundantes,
- los métodos de clasificación de locutores con un mayor poder discriminativo,
- la creación de modelos que reflejen la dinámica del habla,
- los métodos de normalización de rasgos y de resultados de comparación más robustos y efectivos.

Los autores se proponen continuar profundizando en algunas de dichas áreas y, en posteriores reportes de investigación, presentar nuevos resultados.

Referencias

1. Gray, C.H. y Kopp, G.A.: Voiceprint Identification. Bell Telephone Laboratories Report, Bell Laboratories, 1944.
2. Atal, B.S.: Efectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer. 55(6), pp. 1304-1312, 1974.
3. Rosenberg, A.: Listener performance in speaker verification tasks, IEEE Transactions of Audio and Electroacoustics, AU-21: 221-225, 1973.
4. Wolf, J.: Efficient acoustic parameters for speaker recognition. Journal of the Acoustical Society of America, 51 pp. 2044-2056, 1972.
5. Su, L. y Fu, K.: Automatic Speaker Identification using nasal spectra and nasal coarticulation as acoustic clues. Informe del Purdue University School of Electrical Engineering, TR-EE73-33. Air Force Office of Scientific Research Grant. Purdue University, Lafayette, IN, 1973.
6. Li, K. y Hughes, G.: Talker differences as they appear in correlation matrices of continuous speech spectra Journal of the Acoustical Society of America 55: 883-887, 1974.
7. Hollien, H.: Status report on Voiceprint identification in the United States Proceedings of the Carnahan Crime Countermeasures Conference, Oxford U.K. pp. 9-20, 1977.
8. Furui, S.: An overview of Speaker Recognition Technology. ESCA Workshop on Automatic Speaker Recognition, pp. 1-10, 1994.
9. Calvo, J.R.:La identificación automática del locutor con fines criminalísticos. CIDT 2005.
10. Kinnunen, T.:Spectral Features for Automatic Text-Independent Speaker Recognition. Department of Computer Science, University of Joensuu, Finland, 2003.
11. Schafer, R.: A survey of digital speech processing techniques. IEEE Transactions on Audio and Electroacoustics, Vol. 20, Issue 1, pp. 28-35, (1972).
12. Damper, R. y Higgins, J.: Improving speaker identification in noise by subband processing and decision fusion. Pattern Recognition Letters 24, pp. 2167-2173, 2003.
13. Miyajima, C., Watanabe, H., Kitamura, T. y Katagiri, S.: Speaker Recognition Based on Discriminative Feature Extraction – Optimization of Mel-Cepstral Features Using Second-Order All-Pass Warping Function. In Proc. 6th European Conference on Speech Communication and Technology, Eurospeech, Budapest, Hungary, pp. 779-782, 1999.
14. Orman, D. y Arslan, L.: Frequency analysis of speaker identification. In Proc. Speaker Odyssey: the Speaker Recognition Workshop Odyssey, pp. 219-222, Crete, Greece, 2001.
15. Kinnunen, T.: Designing a speaker-discriminative filter bank for speaker recognition. In Proc. Int. Conf. on Spoken Language Processing, ICSLP, pp. 2325-2328, Denver, Colorado, USA, 2002.
16. Besacier, L. y Bonastre, J.-F., *Subband architecture for automatic speaker recognition*, Signal Processing 80, pp. 1245-1259, (2000).
17. Ramachandran, R., Farrell, K., Ramachandran R. y Mammone R.: Speaker recognition - general classifier approaches and data fusion methods. Pattern Recognition 35, pp. 2801-2821, 2002.
18. Higgins, J. E., Damper, R. I. y Harris, C. J.: A Multi-Spectral Data Fusion Approach to Speaker Recognition. Image Speech and Intelligent Systems Reserch Group, Dep. of Electr. and Computer Science, Univ. of Southampton, 1998.
19. Atal, B.S. y Schroeder, M.R.: Predictive Coding of Speech Signals. Report of the 6th Int. Congress on Acoustics, Tokyo, Japan, 1968.
20. Markel, J.D., and Gray Jr., A.H.: Linear Prediction of Speech. Springer Verlag, Berlín, 1976.
21. Huang, X., Acero, A. y Hon, H.-W.: Spoken Language Processing: a Guide to Theory, Algorithm, and System Development, Prentice-Hall, 2001.
22. Makhoul, J.: Linear prediction: a tutorial review. Proceedings of the IEEE 64, 4, pp. 561-580, (1975).
23. Rabiner, L. y Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall, 1993.
24. Rabiner, L. y Schafer, R.: Digital Processing of Speech Signals. Signal Processing, A. Oppenheim, Series Ed. Englewood Cliffs, NJ: Prentice-Hall, 1978.

25. Atal, B.S.: Automatic recognition of speakers from their voices. Proceedings of the IEEE 64(4), pp. 460-475, 1976.
26. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing 29(2), pp. 254-272, 1981.
27. Rosenberg, A. E. y Soong, F. K.: Recent research in automatic speaker recognition. In Advances in Speech Signal Processing, S. Furui and M. M. Sondhi, (eds.) New York: Marcel Dekker, pp. 701-738, 1992.
28. Cole, R. (ed), Mariani, J., Uszkoreit, H., Batista, G., Zaenen, A., Zampolli, A. y Zue, V.: Survey of the State of the Art in Human Language Technology. Cambridge University Press and Giardini, 1997.
29. Campbell, J.P.: Speaker Recognition: A Tutorial. Proceedings of the IEEE, Vol. 85, No. 9, pp. 1437-1462, 1997.
30. Itakura, F.: Line spectrum representation of linear predictive coefficients. Trans. Committee Speech Research, Acoustical Soc. Japan, Vol. S75, p. 34., 1975.
31. Furui, S.: Digital Speech Processing, Synthesis, and Recognition. Second ed. Marcel Dekker, Inc., New York, 2001.
32. Liu, C.-S., Huang, C.-S., Lin, M.-T., y Wang, H.-C.: Automatic speaker recognition based upon various distances of LSP frequencies. In Proc. 25th Annual IEEE International Carnahan Conference on Security Technology, pp. 104-109, 1991.
33. Zilca, R. y Bistritz, Y.: Speaker identification using LSP codebook models and linear discriminant functions. In Proc. 6th European Conference on Speech Communication and Technology, Eurospeech, Budapest, Hungary, pp. 799-802, 1999.
34. Hermansky, H.: Perceptual linear prediction (PLP) analysis for speech. Journal of the Acoustic Society of America 87, pp. 1738-1752, 1990.
35. Openshaw, J., Sun, Z. y Mason, J.: A comparison of composite features under degraded speech in speaker recognition. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Minneapolis, Minnesota, USA, pp. 27-30, 1993.
36. Reynolds, D.: Experimental evaluation of features for robust speaker identification. IEEE Trans. on Speech and Audio Processing 2, pp. 639-643, 1994.
37. Vuuren, S.: Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch. In Proc. Int. Conf. on Spoken Language Processing ICSLP, Philadelphia, Pennsylvania, USA, pp. 1788-1791, 1996.
38. Schroeder, M.R. y Atal, B.S.: Generalized Short-Time Power Spectra and Autocorrelation. Journal of the Acoustical Society of America, 34 (Nov), pp. 1679-1683, 1962.
39. Davis, S.B. y Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. ASSP-28, n 4, Aug., pp. 357-366, 1980.
40. Deller, J. R., Proakis, J. G., y Hansen, J. H.: Discrete-Time Processing of Speech Signals. Prentice Hall, 1993.
41. Reynolds, D. A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, Vol. 17, pp. 91-108, (August), 1995.
42. Hume, J.: Wavelet-like regression features in the cepstral domain for speaker recognition. In Proc. 5th European Conference on Speech Communication and Technology, pp. 2339-2342, Eurospeech, 1997.
43. Soong, F. y Rosenberg, A.: On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. on Acoustics, Speech and Signal Processing 36, 6, pp. 871-879, 1988.
44. Calvo, J. R., Fernández, R. y Hernández, G.: Channel/Handset Mismatch Evaluation in a Biometric Speaker Verification Using Shifted Delta Cepstral Features. L. Rueda, D. Mery y J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 96-105, (2007).
45. Calvo, J. R., Fernández, R. y Hernández, G.: Application of Shifted Delta Cepstral Features in Speaker Verification. Interspeech 2007, August 27-31, Antwerp, Belgium, (2007).
46. Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D. y Deller, J.: Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features. International Conference on Spoken Language Processing, p. 89-92, 2002.
47. Allen, F.: Automatic Language Identification. Ph. Lic., Univ. of New South Wales, Sydney, Australia, 2005.
48. Siafarikas, M., Ganchev, T., Fakotakis, N. y Kokkinakis, G.: Overlapping Wavelet Packet Features for Speaker Verification. Proc of Interspeech , pp. 3121-3124, 2005.

49. Tufekci, Z. y Gowdy, J.N.: Feature extraction using discrete wavelet transform for speech recognition. Proc. of the IEEE SoutheastCon, USA, pp. 116-123, 2000.
50. Long, C.J., Datta, S.: Wavelet based feature extraction for phoneme recognition. Proc. of the ICSLP-96, USA, Vol. 1, pp. 264-267, 1996.
51. Sarikaya, R. y Hansen, H.L.: High resolution speech feature parameterization for monophone based stressed speech recognition. IEEE Signal Proc. Letters, Vol. 7, No. 7, pp. 182-185, 2000.
52. Erzin, E., Cetin, A.E. y Yardimci, Y.: Subband analysis for speech recognition in the presence of car noise. Proc. of the ICASSP-95, USA, Vol. 1, pp. 417-420, 1995.
53. Sarikaya, R. Pellom, B.L. y Hansen, J.H.L.: Wavelet packet transform features with application to speaker identification. Proc. of the IEEE Nordic Signal Processing Symposium, NORSIG'98, Denmark, pp. 81-84, 1998.
54. Farooq, O. y Datta, S.: Mel-scaled wavelet filter based features for noisy unvoiced phonem recognition. Proc. of ICSLP, USA, pp. 1017-1020, 2002.
55. Reyes-Herrera, A.L., Villaseñor-Pineda, L. y Montes-y-Gómez, M.: Automatic Language Identification using Wavelets, Proc. of Interspeech, USA, pp. 401-404, 2006.
56. Siafarikas, M., Ganchev, T. y Fakotakis, N.: Wavelet Packet Based Speaker Verification. Proc. of Odyssey 2004, Spain, 2004.
57. Farahani, F., Georgiou, P. G. y Narayanan, S.: Speaker Identification Using Supra-Segmental Pitch Pattern Dynamics. Proceedings of ICASSP, Montreal, Canada, 2004.
58. Chen. J., Dai, B. y Sun, J.: Prosodic Features Based on Wavelet Analysis for Speaker Verification. Proc. of Interspeech, pp. 3093-3096, 2005.
59. Zheng N., Wang N., Lee T. y Ching, P. C.: Speaker Verification Using Complementary Information from Vocal Source and Vocal Tract. Lecture Notes in Computer Scie nce, Springer, 2006.
60. Goswami, J. y Chan, A. K.: Fundamentals of Wavelets: Theory, Algorithms, and Applications. Wiley, 1999.
61. Mallat, S.: A wavelet tour of signal processing. Academic Press, San Diego, USA, 1998.
62. Battle, G.: A block spin construction of ondelettes. Part I: Lemarié functions. Comm. Math. Phys., 110:601-615, 1987.
63. Lemarié, P. G.: Ondelettes à localisation exponentielle. J. Math. Pures et Appl., 67927-236, 1988.
64. Fernández, R., Montalvo, A., Calvo, J. R. y Hernández, G.: Selection of the Best Wavelet Packet Nodes Based on Mutual Information for Speaker Identification. A ser publicado en CIARP, (2008).
65. Hermansky, H.: RASTA processing of speech. IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, 1994.
66. Hirsch, H. G., Meyer, P. y Ruehl, H. W.: Improved speech recognition using high-pass filtering of subband envelopes. In Eurospeech, Proceedings of the Second European Conference on Speech Communication and Technology, pp. 413-416, Genova, Italy, 1991.
67. Vuuren, S. y Hermansky, H.: On the importance of components of the modulation spectrum for speaker verification. In Proc. Int. Conf. on Spoken Language Processing ICSLP, Sydney, Australia, pp. 3205-3208, 1998.
68. Wet, F.: Additive Background Noise as a Source of non-Linear Mismatch in the Cepstral and Log-Energy Domain. Computer Speech and Language, Vol. 19, pp. 31-54, (2005).
69. Ortega, J., Gonzalez, J. y Marrero, V.: AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech communication (31), pp. 255-264, 2000.
70. Pelecanos, J. y Sridharan, S.: Feature warping for robust speaker verification. In ODYSSEY, pp. 213-218, 2001.
71. Xiang, B., Chaudhari, U. V., Navratil, J., Ramaswamy, G. N., y Gopinat, R. A.: Short-time gaussianization for robust speaker verification. In ICASSP, Vol. 1, pp. 681-684, 2002.
72. Mason, M., Vogt, R., Baker, B., y Sridharan, S.: Data-driven clustering for blind feature mapping in speaker verification. In INTERSPEECH, pp. 3109-3112, 2005.
73. Kenny, P., Boulianne, G., Ouellet, P., y Dumouchel, P.: Factor analysis simplified. In IEEE Conf. Acoust., Speech, and Signal Proc., Vol. 1, pp. 637-640, Philadelphia, Mar 2005.
74. Kenny, P., Boulianne, G., Ouellet, P., y Dumouchel, P.: Improvements in factor analysis based speaker verification. In IEEE Conf. Acoust., Speech, and Signal Proc., Vol. 1, pp. 113-116, Toulouse, May 2006.
75. Vogt, R., Baker, B., y Sridharan, S.: Modelling session variability in text-independent speaker verification. In INTERSPEECH, pp. 3117-3120, 2005.

76. Solomonoff, A., Campbell, W., y Quillen, C.: Channel compensation for SVM speaker recognition. In ODYSSEY, pp. 57-62, 2004.
77. Solomonoff, A., Campbell, W. M., y Boardman, I.: Advances in channel compensation for SVM speaker recognition. In ICASSP, pp. 629-632, 2005.
78. Hatch, A. O., Kajarekar, S., y Stolcke, A.: Within-class covariance normalization for SVM-bsaed speaker recognition. In ICSLP, pp. 1471-1474, Pittsburgh, PA, Sept 2006.
79. Woo, K., Yang, T., Park, K. y Lee, C.: Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters, Vol. 36, No. 2, pp. 180-181, 2000.
80. Li, Q., Zheng, J., Tsai, A. y Zhou, Q.: Robust endpoint detection and energy normalization for realtime speech and speaker recognition. IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 3, pp. 146-157, 2002.
81. Marzinzik, M. y Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 6, pp. 341-351, 2002.
82. Ramírez, J., Yélamos, P., Górriz, J. M. y Segura, J.C.: Speech/Non-Speech discrimination combining advanced feature extraction and SVM learning, In INTERSPEECH, 2006.
83. Duda, R. O., Hart, P. E. y Stork, D. G.: Pattern Classification. John Wiley and Sons, Inc., 2001.
84. Wang, X. y Paliwal, K.: Generalized MCE Training Algorithm for Feature Dimensionality Reduction. Microelectronic Engineering Research Conference, 2001.
85. Brunzell, H. y Eriksson, J.: Feature Reduction for Classification of Multidimensional Data. Pattern Recognition 33, pp. 1741-1748, 2000.
86. Zigel, Y., y Cohen, A.: Text-Dependent Speaker Verification using Feature Selection with Recognition Related Criterion. Odyssey, 2004.
87. Zamalloa, M., Bordel, G., Rodríguez, L. J. y Peñagarikano, M.: Feature Selection based on genetic algorithms for speaker recognition. IEEE Odyssey The Speaker and Language Recognition Workshop, Puerto Rico, 2006.
88. Zamalloa, M., Bordel, G., Rodríguez, L. J., Penagarikano M. y Uribe J. P.: Using Genetic Algorithms to Weight Acoustic Features for Speaker Recognition. Interspeech, 2006.
89. Pandit, M., Kittler, J. y Matas, J.: Selection of Speaker Independent Feature for a Speaker Verification System. ICPR, 1998.
90. Jain, A. y Zongker, D.: Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 19, No. 2, pp. 153-156, 1997.
91. Sambur, M.R.: Selection of Acoustic Features for Speaker Identification. IEEE, Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, pp. 176-182, (Apr.), 1975.
92. Pandit, M. y Kittler, J.:Feature Selection for a DTW Based Speaker Verification System. Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc., pp. 769-772, 1998.
93. Pudil, P., Ferri F.J., Novovicova, J. y Kittler, J.: Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions. Proc. 12th ICPR Jerusalem, 1994.
94. Cheung R.S. y Eisenstein, B.A.: Feature Selection via Dynamic Programming for Text-Independent Speaker Identification. IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-26, No. 5, pp. 397-403, (October) 1978.
95. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Second ed. Academic Press, London, 1990.
96. Hyvärinen, A., Karhunen, J., y Oja, E.: Independent Component Analysis. John Wiley & Sons, Inc., New York, 2001.
97. Jolliffe, I. T.: Principal Component Analysis. Second Edition, Springer, 2002.
98. Tou, J. y Gonzalez, R.: Pattern recognition principles. In Applied Mathematics and Computation, R. Kalaba Ed. Reading, MA: Addison-Wesley, 1974.
99. Woojay, J. y Biing-Hwang J.: A Category-Dependent Feature Selection Method for Speech Signals. Interspech, pp. 365-368, 2005.
100. Hastie, T., Tibshirani, R. y Friedman, J.: The Elements of Statistical Learning. Springer-Verlag, 2001.
101. Sun, Z. P., y Mason, J. S.: Order Analysis of Combined Features in Speaker Recognition. ICSP Proceedings, (1993).
102. Jin, Q. y Waibel A.: Application of LDA to Speaker Recognition. Proc. of the ICSLP, Beijing, China, (October) 2000.

103. Sun, Z. P., Mason, J. S.: Combining features via LDA in speaker recognition. 1993.
104. Zilca, R. D. y Bistritz, Y.: Feature concatenation for Speaker identification. In EUPSICO, 2000.
105. Konig, Y., Heck, L., Weintraub, M. y Sonmez, K.: Nonlinear Discriminant Feature Extraction for Robust text independent speaker recognition. Proc. of RLA2C, Speaker Recognition and its Commercial and Forensic Applications, Avignon, France, 1998.
106. Şentürk, A. y Gürgen, F. S.: Feature Selection by Independent Component Analysis for Robust Speaker Verification. International Journal of Computer Science and Network Security IJCSNS, Vol. 6, No. 3B, March, pp. 229-239, 2006.
107. Holland, J.H.: Adaptation in natural and artificial systems. University of Michigan Press, reprinted in 1992 by MIT Press, Cambridge, MA, 1975.
108. Charbuillet, C., Gas, B., Chetouani, M. y Zarader, J. L.: A New Approach for Speech Feature Extraction Based on Genetic Algorithms. Workshop of Non Linear Speech Processing, 2005.
109. Eriksson, T., Kim, S., Kang, H.-G. y Lee, Ch.: An Information-Theoretic Perspective on Feature Selection in Speaker Recognition. IEEE Signal Processing Letters, Vol. 12, No. 7, 2005.
110. Grall-Maes, E. y Beauseroy, P.: Mutual information-based feature extraction on the time-frequency plane. in IEEE Trans. Signal Process., Vol. 50, pp. 779-790, 2002.
111. Torkkola, K. y Campbell, W. M.: Mutual information in learning feature transformations. In Proc. Int. Conf. Mach. Learning, pp. 1015-1022, 2000.
112. Torkkola, K.: Nonlinear feature transforms using maximum mutual information. In Proc. Int. Conf. Neural Net., pp. 2756-2761, 2001.
113. Ellis, D. y Bilmes, J. A.: Using mutual information to design feature combinations. In Proc. ICSLP, pp. 79-82, 2000.
114. Kwak, N. y Choi, C.-H.: Input feature selection by mutual information based on Parzen windows. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 24, No. 12, pp. 1667-1671, 2002.
115. Hall, M.A. y Smith, L.A.: Practical feature subset selection for machine learning. In Proceedings of the 21st Australian Computer Science Conference, pp. 181-191, 1998.
116. Sakoe, H. y Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26 (1), pp. 43-49, 1978.
117. Soong F.K., et al.: A vector Quantization Approach to Speaker Recognition. AT&T, Tech. J., vol. 66, pp. 14-26, 1987.
118. Linde, Y., Buzo, A. y Gray, R.: An algorithm for vector quantizer design, IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.
119. Jialong, H. y Li, L.: A Discriminative Training Algorithm for VQ-Based Speaker Identification. IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 3, May 1999.
120. Kohonen, T.: The self-organizing map. Proc. IEEE, Vol. 78, pp. 1464-1480, 1990.
121. Kinnunen, T., Kilpelainen, T., y Franti, P.: Comparison of clustering algorithms in speaker identification. In Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000) (Marbella, Spain), pp. 222-227, 2000.
122. Makhoul, S. y Gish, H.: Vector quantization for speech coding. IEEE, Vol. 73, pp. 1551-1587, Nov. 1985.
123. Wang, R.-H., He, L.-S. y Fujisaki, H.: A weighted distance measure based on the fine structure of feature space: application to speaker recognition. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Albuquerque, New Mexico, USA, pp. 273-276, 1990.
124. Matsui, T., y Furui, S.: A text-independent speaker recognition method robust against utterance variations. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Toronto, Canada, pp. 377-380, 1991.
125. Fan, N., y Rosca, J.: Enhanced VQ-based algorithms for speech independent speaker identification. In Proc. Audio- and Video-Based Biometric Authentication, AVBPA, Guildford, UK, pp. 470-477, 2003.
126. Morgan, D.B. y Scofield, C.L.: Neural Networks and Speech Processing, Kluwer Academic Publishers, 1991.
127. Llerena, Y., State of the art in Speaker Recognition. Instituto de Engenharia Electrónica e Telemáica de Aveiro, IEETA. Facultad de Informática, Universidad de Ciego de Ávila, UNICA, 2006.
128. Vivaracho, P. C., Moro, Q. I.: Redes Neuronales Artificiales. Reporte Técnico para el proyecto TIC-2000-1669 C0403, Universidad de Valladolid, Abril 2001.
129. Bennani, Y., Gallinari, P.: Neural networks for discrimination and modelization of speakers. Speech Communication 17, pp. 159-175, 1995.

130. Oglesby, J. y Mason, J.S.: Optimization of Neural Models for Speaker Identification. In Proc. of IEEE Intl. Conf. Acoust. Speech and Signal Proc. (ICASSP'90), pp. 261-264, 1990.
131. Hertz, J., Krogh, A. y Palmer, R.J.: Introduction to the Theory of Neural Computation. Addison-Wesley, Reading, MA, 1991.
132. Rudasi, L. y Zahorian, S.A.: Text-independent talker identification with neural networks, Proc. Internat. Conf: Acoust. Speech Signal Process, S6.6, Toronto, Canada, pp. 389-392, 1991.
133. Oglesby, J. y Mason, J.S.: Radial Basis Function Networks for Speaker Recognition, In Proc. of IEEE Intl. Conf. Acoust. Speech and Signal Proc. (ICASSP'91), pp. 393-396, 1991.
134. Nirajan, M. y Fallside, F.: Neural networks and radial basis functions in classifying static speech patterns, Computer Speech and Language, Vol. 4, pp. 275-289, 1990.
135. Bennani, Y., Fogelman, F. y Gallinari, P.: A Connectionist Approach for Speaker Identification. In Proc. of IEEE Intl. Conf. Acoust. Speech and Signal Proc. (ICASSP'90), pp. 265-268, 1990.
136. Bennani, Y.: Text-independent talker identification system combining connectionist and conventional models. Proc. IEEE Workshop on Neural Networks for Signal Processing, 31 August - 2 September, Copenhagen, Denmark, 1992.
137. Robinson, T.: A real time recurrent error back propagation network word recognition system, Proc. Int. Conf. Acoust. Speech Signal Process, San Francisco USA, pp. 617-620, 1992.
138. Artieres, T. y Gallinari, P.: Neural models for extracting speaker characteristics in speech modelization systems. Proc. Eurospeech, Berlin, Germany, 1993.
139. Hattori, H.: Text-independent speaker recognition using neural networks. Proc. Int. Conf Acoust. Speech Signal Process, Vol. 22, San Francisco, USA, pp. 153-156, 1992.
140. Juang, B. H. y Rabiner, L. R.: Spectral representations for speech recognition by neural networks, a tutorial. Proceedings of the 1992 IEEE-SP Workshop, pp. 214-222, Sept. 1992.
141. Mammone, R., Farell, K. and Assaleh, K.: Speaker recognition using neural networks and conventional classifiers. IEEE Transactions on Speech and Audio Processing, 2(1):194-205, 1994.
142. Ganchev, T., Tasoulis, D.K., Vrahatis, M.N., Fakotakis, N.: Generalized Locally Recurrent Probabilistic Neural Networks for Text-Independent Speaker Verification. Proc. of the ICASSP, Montreal, Quebec, Canada, May 17-21, Vol.1, pp. 41-44, 2004.
143. Doddington, et al.: The NIST speaker recognition evaluation: overview, methodology, systems, results, perspective. Speech Communication, Vol. 31, pp.225-254, 2000.
144. Ortega, J.: Técnicas de mejora de voz aplicadas a sistemas de reconocimiento de locutores. Tesis doctoral, ETSI.Telecomunicación, Universidad Politécnica de Madrid, 1996.
145. Ortega, J. y González, J.: Estudio comparativo de técnicas de identificación automática de locutores. X Symposium nacional de la Unión Científica Internacional de Radio URSI, Valladolid, 1995.
146. Vivaracho, C.E., Ortega, J. y Romero, L.A.: Perceptrón Multicapa frente a Modelos de Mezcla de Gausianas en verificación automática de locutores. Actas del I Congreso de la SEAF, pp.85-90, 2000.
147. Suzuki, T., Tanimoto, M., Osanai, T., Kido, H. y Kamada, T.: Speaker retrieval system for investigative operation. Paper presented in the 82th Annual I.A.I. Conference, Danvers, Massachusetts, 1997.
148. Sistema SVL: Sistema de verificación/identificación de locutores SVL. Documento presentado al laboratorio de Acústica Forense de la D.G.P., 1998.
149. Reynolds, D. y Rose, R.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Vol. 3, No. 1, 1995.
150. Gauvain, J. L. and Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2, pp. 291-298, 1994.
151. Reynolds, D. A.: An Overview of Automatic Speaker Recognition Technology. In ICASSP, pp. 4072-4075, 2002.
152. Reynolds, D. A.: A Gaussian mixture modeling approach to text in dependent speaker identification. Ph.D. thesis, Georgia Inst. of Technology, Sept., 1992.
153. Vuuren, S.: Speaker Verification in a Time-Feature Space. Ph.D. thesis. Oregon Graduate Institute, March 1999.
154. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, pp. 19-41, 2000.

155. Reynolds, D. A.: Comparison of background normalization methods for text-independent speaker verification. In Proceedings of the European Conference on Speech Communication and Technology, pp. 963-966, September 1997.
156. NIST speaker recognition evaluation plans, Philadelphia, PA. Website: www.nist.gov/speech/test.htm.
157. Bimbot, F., Bonastre, J.-F., et al: A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing 4, pp. 430-451, 2004.
158. Martin, A. y Przybocki, M.: The NIST Speaker Recognition Evaluation Series. National Institute of Standards and Technology [Online]. Available: http://www.nist.gov/speech/tests/spk, 2004.
159. Campbell, W., Campbell, J., Reynolds, D., Singer, E., and Torres-Carrasquillo, P.: Support vector machines for speaker and language recognition. Comput. Speech Lang., Vol. 20, pp. 210-229, 2006.
160. Campbell, W., Sturim, D., Reynolds, D., and Solomonoff, A.: SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In Proc. ICASSP, Vol. 1, 2006.
161. Kenny, P. and Demouchel, P.: Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process., Vol. 13, No. 3, pp. 345-354, May 2005.
162. Mariéthoz, J. and Bengio, S.: A Kernel trick for sequences applied to text-independent speaker verification systems. IDIAP, IDIAP-RR 77, 2005.
163. Wan, V.: Speaker verification using support vector machines. Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 2003.
164. Campbell, W. M., Sturim, D. and Reynolds, D. A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett., Vol. 13, No. 5, pp. 308-311, May 2006.
165. Ben, M., Betser, M., Bimbot, F. and Gravier, G.: Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In Proc. ICSLP, pp. 2329-2332, 2004.
166. Scheffer, N. and Bonastre, J.-F.: A UBM-GMM driven discriminative approach for speaker verification. In Proc. Odyssey, pp. 1-7, 2006.
167. Fauve, B. G. B., Matrouf, D., Scheffer, N., Bonastre, J.-F. and Mason, J. S. D.: State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, 2007.
168. Przybocki, M., Martin, A. and Le, A.: NIST speaker recognition evaluation chronicles-Part 2. In Odyssey Workshop Presentation [Online]. Available: http://www.speakerodyssey.com/templates/13.pdf, 2006.
169. Fauve, B. G. B., Matrouf, D., Scheffer, N., Bonastre, J.-F. y Mason, J. S. D.: State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, Sept. 2007.

Anexos

A. Principales instituciones, grupos e investigadores que trabajan en la temática

- Universidad Autónoma de Madrid. Tratamiento de Voz y señales. Investigadores: Javier Ortega, Joaquín Gonzalez y Julian Fierrez-Aguilar.
  `http://atvs.ii.uam.es/`
  `{javier.ortega,joaquin.gonzalez,julian.fierrez}@uam.es`
- AT & T. Speech and Image Processing Services Research Lab. Investigadores: Lawrence R. Rabiner y Aaron E. Rosenberg.
- Bell Laboratories, Lucent Technologies. Multimedia Communication Research labs. Investigadores: Chin Hui-Lee.
- Carnegie Mellon University. CMU Speech Group. Investigadores: Richard M. Stern, Raj Reddy, Evandro B. Gouvêa, Simon Lucey y Qin Jin.
  `{rms,reddy,egouvea,slucey,qjin}@cs.cmu.edu`

http://www.speech.cs.cmu.edu/
http://www.cs.cmu.edu/

- University of Colorado - Boulder. Center for Spoken Language Research. Invesigador: John Hansen.
  John.Hansen@colorado.edu
  John.Hansen@utdallas.edu
  http://cslr.colorado.edu/
- Oregon Graduate Institute. Center for Spoken Language Understanding. Investigadores: Sarel van Vuuren, Hynek Hermansky y Andre Adami.
  http://www.ogi.edu/
  http://cslu.cse.ogi.edu/
  hynek@ece.ogi.edu
  {hynek,adami}@asp.ogi.edu
- Swiss Federal Institute of Technology. Speech Processing & Biometrics Group. Investigadores: Andrzej Drygajlo, Anil Alexander y Filippo Botti.
  http://www.epfl.ch/
  http://scgwww.epfl.ch/
  {andrzej.drygajlo,alexander.anil,filippo.botti}@epfl.ch
- International Computer Science Institute. Speech Group. Investigadores: Kofi Boakye y Barbara Peskin.
  {fkaboakye,barbarag}@icsi.berkeley.edu
  http://www.icsi.berkeley.edu/
  http://www.icsi.berkeley.edu/Speech/
- University of Cambridge. Speech Research Group. Investigadores: Steve Young y Mark Gales.
  {sjy,mjfg}@eng.cam.ac.uk
  http://www.cam.ac.uk/
  http://www.eng.cam.ac.uk/
- Radboud University of Nijmegen, The Netherlands. Automatic Acoustic Recognition Technologies. Investigadores: Johan Koolwaaij, Lou Boves.
  {koolwaaij,boves}@let.kun.nl
  http://www.ru.nl/
- IBM TJ Watson Research Labs. Human Language Technologies. Investigadores: Jason Pelecanos y Jiri Navratil.
  http://www.watson.ibm.com/
  {jwpeleca,jiri}@us.ibm.com
- MIT Massachusetts Institute of Technology. Information Systems Technology Lincoln Laboratory. Investigadores: Douglas Sturim, Thomas F. Quatieri, Pedro T. Carrasquillo, Douglas Reynolds, Douglas A. Jones, Robert B. Dunn, William M. Campbell y Joseph. P. Campbell.
  http://www.ll.mit.edu/IST/
  {ptorres,dar,daj,rbd,wcampbell,jpc}@ll.mit.edu
- SRI Internacional. Speech Technology and Research Lab. “STAR Labs”. Elizabeth Shriberg, Luciana Ferrer, Larry Heck y Francoise Beaufays.
  http://www.sri.com/

http://www-speech.sri.com/
{ees,lferrer,heck,francois}@speech.sri.com

- University of Joensuu. PUMS Project, Computer Science Dept. Investigadores: E. Karpov, T. Kinnunen y P. Fränti.
http://www.joensuu.fi/
http://cs.joensuu.fi/
{eugeny.karpov,tomi.kinnunen,pasi.franti}@cs.joensuu.fi
- University of Avignon. Laboratorie Informatique Avignon LIA. Investigadores: Corinne Fredouille y Jean-François Bonastre.
http://www.univ-avignon.fr/
http://www.lia.univ-avignon.fr
{corinne.fredouille,jean-francois.bonastre}@lia.univ-avignon.fr
- University of Patras, Greece. Wire Communications Laboratory. Investigadores: Nikos Fakotakis, Mihalis Siafarikas y Todor Ganchev.
http://www.upatras.gr
{fakotaki,msiaf,tganchev}@wcl.ee.upatras.gr
- National Chiao-Tung University. Department of Computer and Information Science. Investigador: Hung-Ju Huang.
hungju@cis.nctu.edu.tw
http://www.nctu.edu.tw/
http://www.cis.nctu.edu.tw/
- National Sun Yat-Sen University, Taiwan. Department of Electrical Engineering. Investigador: Chih-Chien Thomas Chen.
chenc@mail.ee.nsysu.edu.tw
http://www.nsysu.edu.tw
- Shanghai University. School of Communication and Information Engineering. Investigador: Chao Wang Li Ming Hou.
lmhou@staff.shu.edu.cn
http://www.shu.edu.cn/
- School of Electronic Engineering, Soongsil University, Seoul. Biometric Engineering Research Center. Investigadores: Jong Joo Lee, Ki Yong Lee.
{jjlee,kylee}@ctsp.ssu.ac.kr
http://www.ssu.ac.kr
- University of Wales, Swansea, UK. Speech Research Group, Department of Electrical Engineering. Investigadores: Kin Yu, John Mason, John Oglesby.
{ky.u,eemasonj}@swansea.ac.uk
- Zhejiang University, P.R. China. College of Computer Science and Technology. Investigadores: Zhenchun Lei, Yingchun Yang, Zhaohui Wu, Pu Yang.
{leizhch,yyc,wzh,kuvun}@zju.edu.cn
http://www.zju.edu.cn
- Universidad Complutense de Madrid. Departamento de Comunicación Audiovisual y Publicidad II. Investigadores: Carlos Delgado Romero, Francisco García García.
cap2@ccinf.ucm.es
fghenche@terra.es

- SRI International. Speech Technology and Research Laboratory. Investigadores: Larry P. Heck, Mitchel Weintraub.
  heck@nuance.co kemal@speech.sri.com
- Universidad Central de Las Villas "Martha Abreu". CEETI. Investigadores: Dr. Juan Lorenzo, Ginori, Dr. Julian Cárdenas Barrera, Alberto Taboada Crispí, Carlos A. Ferrer Riesgo, Rubén Orozco Morales, María E. Hernández-Díaz Huici.
  {juanl,julian,cferrer,mariae,ataboada,rorozco}@uclv.edu.cu

B. Principales revistas, libros y softwares relacionados con la temática

B.1. Revistas

- Speech Communication.
  www.elsevier.nl/locate/specom
- Journal of the Acoustical Society of America.
  http://asa.aip.org/jasa.html
- IEEE Transactions on Speech and Audio Processing.
  http://www.ieee.org/portal/pages/pubs/transactions/tsap.html
- IEEE Transactions on Audio, Speech & Language Processing.
  http://www.ewh.ieee.org/soc/sps/tap/index.html
- Computer, Speech & Language.
  http://www.sciencedirect.com/science/journal/08852308
- Digital Signal Processing.
  http://www.idealibrary.com/links/toc/dspr/10/1/0
- Pattern Recognition.
  http://www.sciencedirect.com/science/journal/00313203
- Pattern Recognition Letters.
  http://www.sciencedirect.com/science/journal/01678655
- International Journal of Pattern Recognition and Artificial Intelligence.
  http://www.worldscinet.com/ijprai/ijprai.shtml

B.2. Libros

- *Survey of the State of the Art of Human Language Technology*,
  Ronald A. Cole
  http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html,
  \\Crpinfo\Documentacion\Libros
- *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*,
  Xuedong Huang, Alex Acero, Hsiao-Wuen Hon y Xuedong Huang
  \\Crpinfo\Documentacion\Libros
- *Automatic Speech and Speaker Recognition: Advanced Topics*,
  Chin-Hui Lee, Frank K. Soong y Kuldip K. Paliwal
- *Discrete-Time Speech Signal Processing: Principles and Practice*,
  Thomas F. Quatieri

- *Discrete-Time Processing of Speech Signals*,
  John R. Deller, John H. L. Hansen y John G. Proakis
- *The HTK Book* (for HTK Version 3.3),
  Cambridge University Engineering Department
  `\\Crpinfo\Documentacion\Libros`
- *The Scientist and Engineer's Guide to Digital Signal Processing*,
  Steven W. Smith
  `\\Crpinfo\Documentacion\Libros`
- *Pattern Recognition in Speech and Language Processing*,
  Wu Chou, Biing Hwang Juang
  `\\Crpinfo\Documentacion\Libros`
- *Pattern Classification*,
  Richard O. Duda, Peter E. Hart Y David G.Stork
  `\\Crpinfo\Documentacion\Libros`
- *Fundamentals of Speech Recognition*,
  Lawrence Rabiner y Biing Hwang Juang

B.3. Softwares

- CMU Sphinx: Librería de reconocimiento del habla (código abierto),
  `http://cmusphinx.org/`
- CSLU Toolkit: Herramientas para la exploración, del conocimiento y la investigación en la voz y human-computer,
  `http://cslu.cse.ogi.edu/toolkit/`
- LNKnet MIT Lincoln Laboratories Pattern Classification Software. LNKnet es un paquete de software que integra más de 22 redes neuronales, clasificación con máquinas de aprendizaje, métodos de agrupamientos y algoritmos de selección de rasgos, máquinas de soportes vectoriales y sencillos clasificadores bayesianos. Tiene una versión para Linux y una para Windows usando el ambiente de Cygwin, todas estas herramientas están creadas en lenguaje C,
  `http://www.ll.mit.edu/IST/lnknet/`
- Cambridge HTK: Biblioteca de algoritmos de HMM. Librería portable para la construcción y manipulación de los HMM. El ATK es un API en tiempo real para HTK,
  `http://htk.eng.cam.ac.uk/`
  `http://mi.eng.cam.ac.uk/~sjy/software.htm`
- LIA, Alize: Herramientas libres para el reconocimiento del locutor. ALIZE es un proyecto de desarrollo iniciado por el consorcio ELISA. El ALIZE presenta una documentación muy bien detallada para su uso, también tiene las siguientes características:
  - es simple y fácil de entender,
  - tiene un nivel de funcionamiento que corresponde con el estado del arte actual, en términos de error pero también en términos de recursos de cómputo necesarios,
  - facilita el desarrollo de demostraciones en aplicaciones prácticas,
  - esta programado en C++.

  `http://www.lia.univ-avignon.fr/heberges/ALIZE/`

- FASR: es un proyecto conjunto del FBI, el Lincoln Lab. del MIT, el US Air Force Research Lab. y BAE System, el prototipo en el 2002 tiene un error menor del 2 % para las peores condiciones de 30 segundos de voz espontánea por diferentes canales, independiente del texto y del canal, parametrización con coeficientes cepstrales en escala Mel, velocidad y aceleración y basado en modelación GMM.
- ASPIC II: es un proyecto conjunto de la Escuela Politécnica Federal de Lausana y el Instituto de Aplicaciones de Inteligencia Artificial, ambos en Suiza, basado en GMM con 32 mezclas, fue evaluado en el 2004 ante la variabilidad del habla, del ruido, de la adquisición, codificación y transmisión, así como del método de grabación, con resultados muy alentadores.
- Identivox (2000): Proyecto elaborado entre la Universidad Politécnica de Madrid y la Guardia Civil de Madrid, basada en GMM en el reconocimiento del locutor independiente del texto, para uso forense, aplicando el modelo bayesiano para las conclusiones. Parametrización con coeficientes cepstrales en escala Mel, su velocidad y aceleración. Normalización del canal y normalización de la medida de la similaridad.
- TRAWL (2004): es un producto comercializado por el Speech Technology Center de Rusia que procesa una plantilla de voz con los primeros tres formantes y establece una medida de diferencia con 5 modelos de voz. Reporta un 91 % de confiabilidad si las dos voces comparadas duran al menos 96 segundos cada una, 85 % si una voz dura 16 segundos y la otra 96 segundos, 90 % si se usa la misma línea, 82 % si las dos voces duran 16 segundos.
- Netlab: Es una librería de herramientas creada en Matlab que proporciona herramientas para el reconocimiento de patrones utilizando modelos de mezclas gaussianas, y redes neuronales para el uso en adiestramiento, investigación y desarrollo de aplicaciones, también proporciona varios ejemplos de su uso y una buena documentación.
  `http://www.ncrg.aston.ac.uk/netlab/index.php`
- BioID America Inc.: El software de desarrollo SDK 3.0 se puede implementar en cualquier aplicación donde las personas deban ser claramente identificadas. Usa reconocimiento facial, verificación de voz y movimientos de los labios para garantizar una elevada exactitud de reconocimiento así como una máxima protección de acceso.
- Classification/Decision techniques: Dynamic Time Warping (DTW) and Hidden Markov Models.
  `http://www.humanscan.de/products/bioid/bioid30.php`
- Anovea Authentication Technology, Inc: Descripción: La familia de productos Anoveas se basa en un propietario, sigue una estrategia para el procesamiento de señales del multirasgos que ofrece umbrales de errores bajos y una alta invariación (robustez) ante el canal. El sistema permite a los programadores agregar la verificación del locutor en ambientes independientes de servidores y de cliente. Consiste en un equipo de desarrollo de software de SVLib y de SVLite (SDKs) así como también un sistema de AnoveaWeb.
  `http://www.anovea.com/www/products_sap.htm`
- VoiceVault Services: Los servicios de VoiceVault proporcionan el reconocimiento integrado de la voz y la verificación del locutor.
  `http://www.voicevault.com/services.php`

RT_008, febrero 2008
Aprobado por el Consejo Científico CENATAV
Derechos Reservados © CENATAV 2008
Editor: Lic. Miriela Santos Toledo
Diseño de Portada: DCG Matilde Galindo Sánchez
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba