RNPS No. 2142
ISSN 2072-6287
Versión Digital

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

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Métodos para reducir la variabilidad de sesión en el reconocimiento del locutor

Ana Moltalvo Bereau
y José R. Calvo de Lara

RT_051 **noviembre 2012**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
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 para reducir la variabilidad de sesión en el reconocimiento del locutor

Ana Montalvo Bereau
y José R. Calvo de Lara

RT_051 **noviembre 2012**

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

Tabla de contenido

Métodos para reducir la variabilidad de sesión en el reconocimiento del locutor

Ana Montalvo Bereau y José R. Calvo de Lara

Dpto. Reconocimiento de Patrones, Centro de Aplicaciones de Tecnologías de Avanzada(CENATAV),
La Habana, Cuba
{amontalvo, jcalvo}@cenatav.co.cu

RT_051, Serie Azul, CENATAV
Aceptado: 13 de agosto de 2012

Resumen. El presente reporte técnico aborda la problemática de la robustez de los sistemas automáticos de reconocimiento del locutor (SARL) ante la variabilidad de sesión. En el mismo se realiza un análisis crítico de los métodos más importantes reportados en la literatura. Del conjunto de métodos analizados, se seleccionó un subconjunto y apoyados en la herramienta libre ALIZE, se desarrollaron experimentos con fines investigativos y de exploración de la herramienta.

Palabras clave: reconocimiento robusto del locutor, variabilidad de sesión.

Abstract. This report faces the problem of intersession variability in automatic speaker recognition systems. A critical study of the most relevant methods reported in the scientific literature is made. From the set of studied methods, a subset of methods was selected and, based on the free toolkit ALIZE, some experiments were done looking for explore most of ALIZE toolkit potentialities.

Keywords: robust speaker recognition, intersession variability.

1 Introducción

A grandes rasgos, un Sistema Automático de Reconocimiento del Locutor (SARL) actúa como un clasificador de patrones. Cada patrón está formado por un conjunto de características o parámetros, extraídos de una determinada locución y es enfrentado o comparado con distintos modelos generados para cada locutor.

La salida del clasificador ofrece una verosimilitud o una medida de distancia, entre el patrón de entrada y el modelo; y en última instancia una decisión, basada en un umbral, que clasifica la locución como perteneciente a un determinado locutor o a ninguno. Cada modelo de un locutor es generado mediante patrones extraídos de locuciones del mismo; siendo necesario que cada locutor involucrado en el sistema, disponga de su propio conjunto de datos de entrenamiento. Este conjunto será distinto del conjunto de datos sobre los cuales se pruebe el sistema.

Un factor que afecta sensiblemente el rendimiento de los SARL 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. Pero las fallas de estos sistemas no son debidas solamente a las variaciones de canal, ruido ambiente o similaridades entre locutores diferentes,

hay una componente grande al problema que son las diferencias entre dos locuciones hechas por la misma persona, o 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 empleando datos obtenidos bajo condiciones muy distintas.

Es para lidiar con este efecto negativo que se realizan estudios buscando robustecer el sistema, lo que se ha convertido en un tema de investigación importante dentro del área.

En este reporte expondremos distintos métodos empleados para contrarrestar la variabilidad acústica entre los conjuntos de entrenamiento y prueba, veremos que para esto se desarrollan estrategias de compensación a diferentes niveles, que se conocen también por técnicas de normalización.

1.1 Revisión de métodos

Actualmente existen diversas y muy variadas técnicas aplicadas a la compensación o eliminación de la variabilidad de sesión. Estas pueden agruparse en: técnicas de normalización de rasgos, de normalización de modelos y de normalización de la puntuación, en función del nivel sobre el que actúan.

En el grupo de normalización a nivel de rasgos los métodos más utilizados son la Normalización Cepstral de la Media y la Varianza (CMVN por sus siglas en inglés) [2], que ha probado ser muy eficiente para disminuir los efectos lineales del filtrado del canal de transmisión; la Técnica Espectral Relativa (RASTA) [3] y Sustracción Cepstral de la Media (CMS), que fueron introducidas originalmente de manera muy vinculada con el procesamieto de Predicción Perceptual Lineal (PLP) es decir con un filtrado pasa-banda en los dominios logarítmicos o cepstrales, por tanto las variaciones pequeñas de la repuesta del canal serían eliminadas; Compresión de Rasgos o “Feature Warping” [4,5] y Mapeo de rasgos o “Feature Mapping” [18], otras dos propuestas interesantes y muy empleadas para contrarrestar las distorsiones inherentes al canal. Se exponen también al abordar estas temáticas las técnicas de Análisis Discriminativo Lineal (LDA), Análisis Discriminativo No Lineal (NLDA) [20] y Análisis Discriminativo Lineal Heterocedástico (HLDA), ya que mejoran sensiblemente los resultados cuando acompañan la modelación de variabilidad de sesión. Sin embargo son métodos de reducción de dimensionalidad y por no ser este el objetivo central del presente reporte no son explicados.

La compensación a nivel de modelo implica modificar los parámetros del modelo del locutor en lugar de los vectores de rasgos. Estas técnicas buscan contrarrestar la variabilidad de sesión, y han probado ser capaces de compensar las variaciones sin necesidad de explicitar las condiciones en las que fueron obtenidos los datos. La mayoría de estas técnicas se basa en modelar la variabilidad de las frases restringiéndolas a un espacio de menor dimensión. Son exponente de este método el Análisis de Factores (FA) [14], Proyección de los Atributos no Deseados (NAP, Nuisance Atributte Projection) [13], Normalización Intra-Clase de la Covarianza (WCCN, Within class covariance Normalisation) [17] y de gran impacto en el estado del arte, i-vector [30].

Por su parte, la normalización de la puntuación busca eliminar el desplazamiento o corrimiento de las puntuaciones debidas a variaciones en las condiciones del canal. Estos métodos normalizan la verosimilitud logarítmica resultante justo antes de tomar la decisión. En este caso la puntuación es una probabilidad logarítmica generada a partir de un modelo estadístico, y como el cálculo de las mismos (puntuaciones) depende del medio en el que fueron hechas las pruebas, el propósito de esta normalización es reducir la diferencia entre las condiciones en las que fue realizado el entrenamiento y en las que fue realizada la prueba. Las técnicas de normalización a nivel de puntuación comenzaron siendo mayormente utilizadas en la verificación y actualmente son aplicadas también a la identificación, básicamente incluyen a Normalización Cero o Z-norm [23],[24], Normalización Prueba o T-norm y Normalización del auricular o

HNorm [19] y Normalización para celulares o CNorm [18]. Es importante conocer que las técnicas de normalización comprenden también la aproximación o adaptación del clasificador usado [28].

2 Técnicas de normalización a nivel de rasgo

En principio, es posible utilizar una técnica genérica de supresión de ruido para mejorar la calidad de la señal en el espacio temporal, previo a la extracción de rasgos, de hecho es común que se le realice un preprocesamiento a la señal. Sin embargo, es igualmente necesario diseñar un extractor de rasgos que brinde robustez a la representación ante variaciones de sesión en general, así como normalizar los rasgos antes de ser empleados por los algoritmos de modelación [1].

En una parametrización basada en coeficientes cepstrales [1][21], una locución, es dividida en cortas ventanas de tiempo (20*ms*), de la cual son extraídos un cierto número de coeficientes cepstrales. *CMS* se basa en sustraer a cada coeficiente cepstral extraído, la media de dicho coeficiente a lo largo de toda la locución. De esta forma se reduce la distorsión introducida por elementos de variación lenta, como por ejemplo ruido estacionario [22]. *RASTA* explota las diferencias entre las propiedades temporales de la voz y las propiedades temporales de las distintas distorsiones de canal con el objetivo de reducir el efecto del canal de comunicaciones en el espectro. Las características del canal varían poco con el tiempo, por lo que sus componentes espectrales son de baja frecuencia, donde no hay demasiada información de la voz. RASTA filtra en el tiempo los valores de energía en cada banda de frecuencias, con el objetivo de eliminar dicha componente de baja frecuencia [3]. Para eliminar los efectos acústicos no lineales en el dominio de los rasgos, han sido introducidos métodos como *Ecualización de Histogramas* (HEQ) [31], *Normalización Cepstral de Momentos de Mayor Orden* (HOCMN) [32] y muchos otros que no son más que extensiones de las normalizaciones de la media y la varianza pero que normalizan empleando momentos mayores que dos. *Compresión de rasgos* no solo modifica los parámetros estadísticos de los datos como CMS que lo hace modificando la media, sino que actúa también sobre la función de densidad de probabilidad de los mismos para acomodarlos a una distribución normal. De esta forma se puede compensar la variación de canal, el ruido aditivo y hasta cierto punto, efectos no lineales debidos a los transductores . *Mapeo de rasgos* compensa la distorsión de canal estudiando las diferencias de las distribuciones de datos no afectados por el canal (al menos idealmente), y aquellos afectados por un tipo concreto de canal, para aplicar después la transformación inversa (compensar el desplazamiento) a la que produjo el canal [18].

2.1 Normalización cepstral de la media y la varianza (*CMVN)*

El método de normalización a nivel de rasgo CMVN, es uno de los más ampliamente utilizados, es incluso una base para otras normalizaciones [33]. En esta aproximación los rasgos son linealmente transformados de manera tal que resultan rasgos con media cero y varianza unitaria. Para una secuencia temporal de rasgos, la normalización tendrá lugar de la siguiente manera:

$$X = \{x(1), x(2), ..., x(N)\}, \tag{1}$$

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x(n), \qquad \sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N} (x(n) - \overline{x})^2, \tag{2}$$

[1] Conjunto de rasgos de uso más común actualmente en el reconocimiento del habla y del locutor. Su representación en escala logarítmica y distorcionados en frecuencia con escala Mel conforman los conocidos Coeficientes Cepstrales en escala Mel (MFCC)

$$\hat{x}(n) = \frac{x(n) - \overline{x}}{\sigma_x}. \tag{3}$$

Este método de nomalización se basa en asumir que el nivel de ruido es consistentemente estable a lo largo de una alocución dada, por lo que sustraer el vector de media a cada vector de rasgos ayuda a eliminar el ruido de fondo y de canal. Sin embargo hay que señalar que con esta sustracción se pierden características del locutor. Muy similar a la sustracción cepstral se encuentra la sustracción espectral, cuya diferencia radica en que se trabaja en el dominio espectral por lo que comparten ventajas y desventajas.

2.2 Técnica espectral relativa (*RASTA)*

Este procesamiento transforma el espectro logarítmico aplicando un conjunto de filtros pasa-banda con el fin de eliminar o suprimir las componentes constantes o que varíen suavemente, las cuales reflejan el efecto de factores convolucionales en los canales de comunicación. Como es sabido [3] las distorsiones lineales, causadas por los canales de comunicación o diferentes micrófonos, aparecen como una constante aditiva en el espectro logarítmico. Entonces con filtros pasa-banda en el dominio logarítmico, los ruidos aditivos de canal son sustancialmente disminuídos, y esta es la idea central de RASTA. Concretamente RASTA se lleva a cabo en el dominio espectral logarítmico escala bark, descrito por [3], y la función de transferencia del filtro es:

$$H(z) = 0{,}1z^4 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0{,}98z^{-1}}. \tag{4}$$

La correspondiente porción paso-alto del filtro descrito en 4 se espera que atenúe el efecto del ruido convolucional introducido por el canal, mientras que la parte paso-bajo suaviza los cambios espectrales que se producen al pasar de una ventana a otra.

2.3 Ecualización de histogramas (*HEQ)*

De manera general, el ambiente acústico induce una transformación no lineal en el dominio cepstral, y las transformaciones lineales antes expuestas solo pueden eliminar parcialmente este efecto. Es por lo que se discute a continuación un método clásico, que hace uso de aproximaciones mucho más generales en las cuales son usadas transformaciones no lineales. HEQ es introducido en [34] para reducir las diferencias entre las locuciones telefónicas y las grabadas, y más adelante fue exitosamente usado para reconocimiento robusto del habla con ruido. En esta aproximación la función de distribución de probabilidad es usada como una técnica de normalización en la cual los datos de entrenamiento y prueba para cada coeficiente cepstral son aproximados por un histograma. Los histogramas son construidos usando 100 intervalos equiespaciados en el rango $[\mu - 4\sigma, \mu + 4\sigma]$, donde μ y σ son la media y la varianza del coeficiente que será ecualizado, o sea ajustado entre determinados valores.

2.4 Normalización del auricular (*H-norm)*

H-norm es un método de normalización del auricular, por lo que es utilizado para normalizar señales obtenidas a través de distintos canales. Es válido aclarar que el método que exponemos a continuación no es el presentado en [35], sino que es introducido por Wu et al.[27]. Para el rasgo x_i de la i-'esima ventana y su correspondiente energía e_i, las energías están divididas en L niveles $E_l, l \in [1, L]$ y para cada nivel se

calcula un vector de media m_l:

$$m_l = \frac{1}{N_l} \sum_{E_l \leq x_n \leq E_{l+1}} x_n, \tag{5}$$

donde N_l es el número de tramas cuya energía está en el intervalo $[E_l; E_l + 1]$. H-norm pudiera verse como una sustracción cepstral que va restando a los rasgos, distintas medias en función de su nivel de energía correspondiente:

$$\hat{x} = x_i - m_l \quad y \quad e_i \in [E_l; E_{l+1}]. \tag{6}$$

2.5 Normalización para celulares (*C-norm)*

C-norm es referido como un método de normalización para aplicar en celulares, el cual fue propuesto en [18] para combatir los efectos del canal en la telefonía celular. Sin embargo C-norm también es clasificado como un método de mapeo de rasgos, porque se basa en una función que transporta de un espacio de rasgos dependiente del canal (DC) a uno independiente del canal (IC) y el reconocimiento es realizado en este nuevo espacio. Denotemos a x_n como los rasgos de la n-ésima trama en el espacio DC y y_n a aquellos rasgos de la n-ésima trama en el espacio IC. Las mezclas gaussianas que modelan el espacio DC e IC serán G^{DC} y G^{IC} respectivamente. El Modelo de Mezclas Gaussianas (GMM) al cual pertenece x_n es elegido de acuerdo al criterio:

$$i = argmax_j w_j^{DC} \times p_j^{DC}(x(n)|\mu_j^{DC}, \sigma_j^{DC}), \tag{7}$$

donde un modelo de mezclas gaussianas es definido por su peso, media y desviación estándar $\{w_j^{DC}, p_j^{DC}, \sigma_j^{DC}\}$.

De este modo, aplicando la transformación, un vector de rasgos del espacio IC y_n es mapeado a partir de x_n de acuerdo a:

$$y_n = f(x_n) = (x_n - \mu_j^{DC}) \frac{\sigma_j^{IC}}{\sigma_j^{DC}} + \mu_j^{IC}, \tag{8}$$

donde j es la mezcla Gaussiana a la cual x_n pertenece y es determinada en términos de 6. Luego de la transformación, el reconocimiento es llevado a cabo en el espacio IC.

2.6 Mapeo de rasgos o feature mapping

El Mapeo de rasgos 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 [7] se propone una técnica de mapeo de rasgos para lidiar con este inconveniente. Este enfoque parte de la hipótesis de que la distorsión producida por el canal afecta a los diferentes modos de la distribución estadística de los datos mediante un desplazamiento geométrico de los mismos. Bajo esta hipótesis es posible compensar esta distorsión del canal estudiando las diferencias de las distribuciones de datos no afectados por el canal (al menos idealmente), y aquellos afectados por un tipo concreto de canal, para aplicar después la transformación inversa (compensar el desplazamiento) a la que produjo el canal [18].

En la subsección 3.1 se explica el método de Síntesis del Modelo de un Locutor (SMS) cuya esencia es muy similar al Mapeo de Rasgos, solo que mientras SMS se concentra en obtener modelos de locutores por canales nuevos, el Mapeo de Rasgos o Feature Mapping intenta transportar los rasgos provenientes de diversos canales a un espacio común independiente del canal. Ambas aproximaciones están relacionadas

en que las dos aprenden las transformaciones a partir de examinar como los parámetros del modelo cambian y se reescalan luego de la adaptación MAP. El Mapeo de rasgos es motivado por varios factores. En primer lugar, una aproximación en el dominio de los rasgos tiene potencialmente más amplio uso desde el punto de vista de que no está atada a ninguna estructura o modelo particular. En segundo lugar mapear todos los rasgos a un mismo espacio permite acumular información obtenida por diferentes tipos de canales. No obstante ambas técnicas compensan la variabilidad de canal de forma discreta como se expone más adelante.

En el Mapeo de rasgos, al igual que en SMS se entrena un GMM independiente del canal llamado GMM raíz obtenido de datos con variados canales, y se obtienen modelos dependientes del canal de adaptar el GMM raíz. Los parámetros de los modelos cambian de un canal a otro y respecto al modelo independiente del canal, y se observa como están relacionadas las variaciones de las distribuciones en el domino de los rasgos para cada canal, lo que se utiliza para crear la función de mapeo de rasgos.

3 Técnicas de normalización a nivel de modelo

El estado del arte en el reconocimiento del locutor está mayormente dominado por el uso de los Modelos de Mezclas Gaussianas (GMM) y las Máquinas de Soporte Vecotrial (SVM) como clasificadores, incluso también desde una perspectiva de rasgos, con los rasgos cepstrales de corto tiempo [46], [38]. Una amplia variedad de técnicas han sido propuestas para enfrentar la variabilidad de sesión a nivel de modelo entre las que se destacan Síntesis del Modelo del Locutor (SMS), Proyección de atributos indeseables (NAP), Normalización de la Covarianza Intra-clase (WCCN), Análisis de Factores (FA) y una muy recientemente explotada: i-vector (Vector Intermedio).

3.1 Síntesis del modelo del locutor (*SMS)*

SMS es un algoritmo para contrarrestar dentro de la variabilidad de sesión, la diferencia de canal entre los modelos de los locutores entrenados y los locutores de la prueba. En este método de compensación, la modelación busca disminuir los efectos del canal utilizando los parámetros dependientes del canal de los modelos como información o conocimieto previo.

A modo de paréntesis es oportuno definir que dado un segmento de habla X y un locutor hipótesis S, la tarea de 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).

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 la hipótesis alternativa está menos bien definido, puesto que, potencialmente, debe representar el espacio entero de alternativas posibles al locutor supuesto. 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 se entrena para representar la hipótesis alternativa al locutor. Varios términos utilizados para este modelo son: modelo general, modelo del mundo y modelo universal de background (UBM, por sus siglas en inglés).

Introducido en [48] SMS adapta los parámetros del modelo GMM del locutor, a nuevas condiciones del canal, por las cuales no se posee señales para entrenar. Esto se lleva a cabo con la ayuda de trans-

formaciones entre un UBM independiente del canal o UBM Raíz y modelos adaptados dependientes del canal.

La estructura para la construcción de este modelo se muestra en la figura 1:

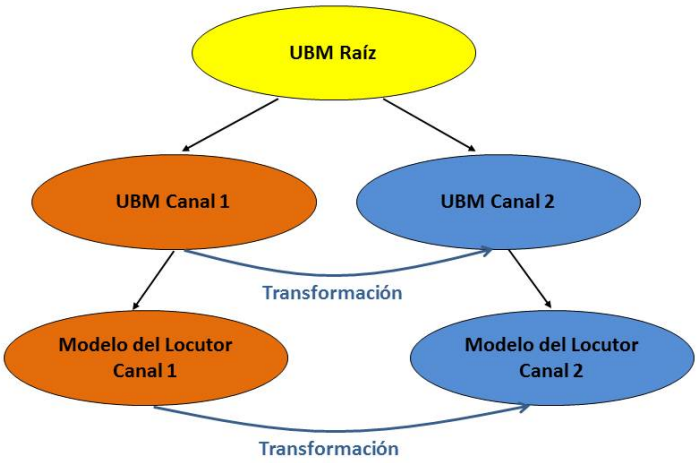

Fig. 1. Estructura para la construcción del modelo sintetizado a partir del UBM deseado.

Para el modelo de un locutor, entrenado por el canal 1, los parámetros del modelo sintetizado en el canal 2 se estiman de la siguiente manera:

$$\mu^{c2}_{loc,i} = \mu^{c1}_{loc,i} + (\mu^{c2}_{ubm,i} - \mu^{c1}_{ubm,i}), \tag{9}$$

$$\omega^{c2}_{loc,i} = \omega^{c2}_{ubm,i}, \tag{10}$$

$$\Sigma^{c2}_{loc,i} = \Sigma^{c2}_{ubm,i}, \tag{11}$$

donde $\mu^{c1}_{loc,i}$ y $\mu^{c1}_{ubm,i}$ son los vectores de media de los modelos de los locutores originales entrenados y el correspondiente UBM dependiente del canal por el canal 1 respectivamente. Por su parte $\mu^{c2}_{ubm,i}, \omega^{c2}_{ubm,i}$ y $\Sigma^{c2}_{ubm,i}$ son los parámetros del UBM dependiente del canal relativos al canal 2. Como los modelos de los locutores son usualmente entrenados con los UBMs dependientes del canal apartir de adaptar solamente los vectores de media, los pesos y varianzas del modelo sintetizado son los del correspondiente UBM dependiente del canal.

SMS aprende cómo cambian los parámetros del modelo del locutor sobre difrentes canales donde no hay datos de entrenamiento disponible, y sintetiza un modelo del locutor para ellos. Este algoritmo asume que todos los locutores están sujetos a las mismas transformaciones al pasar de un canal a otro, sin embargo en la realidad esto no es así.

En [49] buscando hacer dependiente del locutor la transformación se propone un SMS basado en un cohorte. Para este método se requiere un conjunto de locutores, en el que cada locutor tiene un modelo

para cada canal comprendido. Como desventajas del método se pudiera mencionar el fuerte requerimiento de tener un UBM de la mayor cantidad de canales posibles, para poder sintetizar los modelos al canal buscado y la manera discreta de aproximación a la variabilidad de canal.

La naturaleza de la variabilidad de sesión es continua [47], lo que está en contraste con SMS, ya que asume una colección discreta de condiciones de grabación para modelar el fenómeno y contrarrestarlo. Esta técnica de modelación discreta de la variabilidad de sesión impide modelar condiciones del canal que "caigan entre" condiciones que no fueron vistas en el entrenamiento. Es por esto que en el estado del arte se utilizan métodos continuos para modelar la variabilidad de sesión como los que serán expuestos en las siguientes secciones.

3.2 Supervectores (SV)

En los primeros estudios [36] los modelos eran generados por un promedio de los rasgos en el tiempo de manera que cada ventana de 20ms de la secuencia de habla era representada por un único vector. Los vectores promedio eran entonces comparados usando una medida de distancia [37], lo cual es computacionalmente muy eficiente pero brinda poca exactitud en el reconocimiento. Recientemente se ha retomado esa representación que utiliza un único vector, llamado supervector.

Estos supervectores pueden ser usados como entrada de las máquinas de soporte vectorial 2 o de los convencionales modelos del locutor de mezclas gaussianas adaptadas. La combinación de modelos generativos con SVM ha llevado a muy buenos resultados [38].

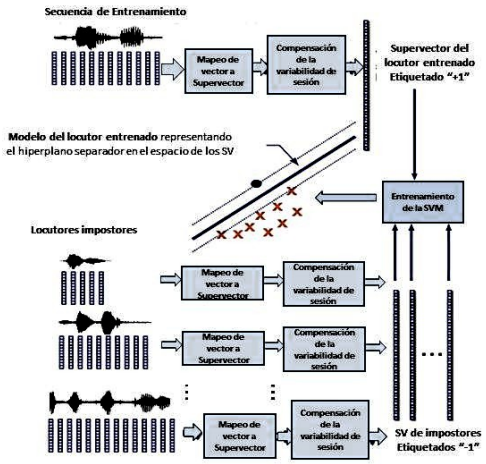

Fig. 2. Vectores de rasgos mapeados al espacio de los supervectores seguidos de una compensación de la variabilidad de sesión y del entrenamiento de la SVM.

Con frecuencia se refieren a "supervectores" como la concatenación de muchos vectores formando uno de mayor dimensión, por ejemplo, unir los vectores de media d-dimensionales de K componentes GMM para obtener un supervector Kd-dimensional [38], no obstante se puede entender el supervector de una manera amplia como una representación de dimensión fija de una secuencia de señal.

Técnicas para compensar la variabilidad de sesión a este nivel han sido desarrolladas recientemente [39], [40], [41]. A partir de que cada secuencia de habla es representada por un único punto en el espacio de los supervectores, se hace prácticamente más alcanzable cuantificar y eliminar la variabilidad indeseada de los supervectores. Cualquier variación de un mismo locutor, caracterizada por sus supervectores y debido a cambios en el canal, en el ambiente o en el contexto fonético, es indeseable.

3.2.1 Análisis de factores

El análisis de factores es un método estadístico usado para describir la variabilidad de variables observables en términos de un menor número de variables ocultas o latentes llamadas factores. Se parte de asumir que un número X de variables observables reflejan las variaciones de un menor número de variables ocultas. FA busca esas variaciones conectadas, como respuesta a las variaciones de las latentes. Las variables observables son modeladas como combinación lineal de las latentes más un término de error.

En una aproximación generativa basada en GMM, cada locutor es representado por un GMM compuesto por C Gaussianas. Estas Gaussianas son estimadas en un espacio de parámetros, continuo, de dimensión F. Cada Gaussiana es caracterizada por un vector de media, una matriz diagonal de covarianza y los pesos. El modelo GMM de un locutor es construido adaptando a sus rasgos las mezclas GMM del UBM el cual es construido a partir de un conjunto grande de locutores. Es en esta adaptación que interviene FA.

Con el objetivo de formular correctamente la teoría tras FA, se asume primeramente que el supervector M, dependiente del canal y del locutor, puede ser representado como una combinación lineal de estos términos, representados por supervectores estadísticamente independientes y normalmente distribuídos,

$$M = s + c. \tag{12}$$

En segundo lugar se asume que tanto la componente del locutor (s) como la del canal (c), pueden ser descritas en función de variables ocultas en la forma:

$$s = m + Vy + Dz, \tag{13}$$

donde m es el supervector de medias del UBM de dimensión $CF \times 1$, V es la matriz rectangular cuyos vectores columna son los *vectores propios del locutor* (eigenvoices [2]), D es una matriz diagonal $CF \times CF$ y z y y son vectores latentes $CF \times 1$ asociados a una distribución normal.

$$c = Ux, \tag{14}$$

donde U es una matriz rectangular cuyos vectores columna son los *vectores propios del canal* (eigenchannel[3]) y x los factores del canal.

La mayoría de los autores utilizan el término FA para referirse solamente al modelado de la contribución del canal (14) y como la técnica tiene un alcance mucho más amplio, cuando se incluye el modelo del locutor (13) es llamado Análisis de Factores Conjunto (JFA).

En el caso particular de $y = 0$, 13 describe exactamente la misma adaptación que la técnica de máximos a posteriori [4] (MAP) [46], por lo que el modelo del locutor de JFA puede ser visto como una extensión a la técnica MAP con el eigenvoice incluído, lo que ha demostrado sus ventajas en el entrenamiento con muestras cortas [47].

[2] Son los autovectores de la matriz de covarianza de un conjunto de supervectores de muchos locutores.

[3] Son los autovectores de la matriz de covarianza de un conjunto de supervectores por diversos canales.

[4] Algoritmo que ha sido clave en las mejoras de los clasificadores, alcanzando el estado del arte durante la evaluación del National Institute for Standars and Technology (NIST, 2004)

El modelo de JFA ([42], [43], [40]) tiene en cuenta la variabilidad del locutor y de la sesión en el contexto de las GMM. Tradicionalmente usado en conjunto con los rasgos cepstrales, JFA ha sido extendido a otros rasgos ([44], [45]). Este modelo se basa en la combinación de la clásica adaptación MAP [40] y eigenvoice para modelar la variabilidad del locutor, unido a la variante MAP eigenchannel que tiene en cuenta la variabilidad de sesión.

Las matrices U, V y D son los llamados hiperparámetros del modelo JFA. Estas matrices son estimadas de antemano sobre grandes conjuntos de datos. Una forma posible de hacerlo es estimar primero V seguida de la estimación de U y D [14], [40]. Para una determinada muestra de entrenamiento, los factores latentes del canal x y del locutor y son estimados de manera conjunta y luego se estima z. Finalmente el supervector del canal c, es descartado y el supervector del locutor s es asumido como el modelo del locutor. De esta forma es realizada la compensación de canal a través del modelado explícito de la componente del canal durante el entrenamiento [47].

Diferentes evaluaciones de diversos grupos de investigación han demostrado el potencial de JFA, sin embargo el método tiene algunas deficiencias prácticas. Una de estas es la sensibilidad, tanto en el entrenamiento como en la prueba, al largo de las señales y sus diferencias, especialmente para señales cortas (10-20s). Los autores de [62] defienden que esta dependencia es mayormente debida a la variabilidad intra-sesión y no a la inter-sesión capturada por la línea base del JFA. Luego, extendieron el modelo JFA añadiendo explícitamente un modelo de variabilidad intra-sesión. Esto probó tener mejores resultados que la línea base JFA cuando el entrenamiento y la prueba son realizados con señales de distintas duración, incluso cuando se desconoce el largo de la señal.

3.2.2 NAP

NAP es un exitoso método para compensar la variabilidad de sesión, muy aplicada a los supervectores SVM [13]. No es específico a un Kernel en particular, por lo que puede ser aplicado a cualquier tipo de supervectores SVM. La transformación NAP elimina las direcciones indeseadas de variabilidad de sesión contenida en los supervectores, antes del entrenamiento de la máquina de soporte vectorial.

La aproximación NAP fue introducida para lidiar con problemas de variabilidad de sesión en el ambiente de SVMs aplicadas a la verificación de locutores en [13], y aplicada con éxito en [58] a un kernel lineal (KL). Este método utiliza una apropiada matriz de proyección, en el espacio del kernel con el objetivo de eliminar allí la variabilidad indeseada (de canal por ejemplo).

Para esta normalización se requiere un conjunto de datos diseñado con gran variabilidad de canal, y los pasos para su estimación de acuerdo a [50], serían:

- Se parte de supervectores de dimensión [5] D_{SV} para cada sesión de cada locutor.
- Para cada locutor se obtiene un supervector de medias sobre todos los SV disponibles del locutor en cuestión, y luego se sustrae este supervector obtenido a cada supervector original. Mezclándose todo estos vectores resultado de la diferencia se obtiene una gran matriz D de supervectores de la cual se ha eliminado la mayoría de la variabilidad del locutor pero que sigue conteniendo la variabilidad de sesión. Las dimensiones de esta matriz $D_{SV} \times N_{ses}$, donde N_{ses} es el número total de sesiones.
- Se seleccionan las dimensiones de la transformación NAP D_{NAP}, la cual por lo general es determinada empíricamente.
- Se realiza un PCA sobre la matriz D, lo que significa encontrar los D_{NAP} vectores propios de la matriz de dispersión normalizada $\frac{1}{N_{ses}} DD'$ (este problema de autovalores podría complicarse computacionalmente debido a las dimensiones que pudieran llegar a tener estas matrices).

[5] Para los supervectores GMM la dimensión coincide con la de los rasgos multiplicada por la cantidad de mezclas gaussianas

- Una vez obtenida la matriz $D_{SV} \times N_{ses}$ que denotaremos E, es una buena precaución normalizar los vectores columna y ortogonalizarlos mutuamente, ya que si los vectores no son ortonormales la transformada NAP falla al proyectar el subespacio fuera.

La proyección NAP: Con la matriz E se entrenan los modelos de la SVM buscando mayor robustez. La básica transformación NAP está diseñada para ser aplicada con SVMs que empleen kernels lineales. La transformación debe aplicarse a todos los supervectores (target y backgrounds) antes de ser utilizados en el entrenamiento del modelo SVM. Esto es que cada supervector v es transformado como:

$$\bar{v} = v - E(E'v). \tag{15}$$

La ortonormalidad garantiza que no sea necesario aplicar la transformada NAP también a los vectores de prueba, además nótese que la transformación se realiza antes de entrenamiento de la SVM, no tiene sentido alguno realizarla a los modelos.

3.2.3 WCCN

WCCN es un método de compensación de supervectores SVM, muy similar a NAP, propuesto en [17]. Se consideran kernels lineales generalizados de la forma $K(s_1, s_2) = s_1 R s_2$, donde s_1 y s_2 son supervectores y R es una matriz semidefinida positiva [6]. Sin perder generalidad, el error de un clasificador binario puede ser minimizado al elegir $R = W^{-1}$, donde W es la esperada matriz de covarianza intralocutor. WCCN fue combinado con PCA en [17] para atacar el problema de estimar e invertir W para conjuntos grandes de datos. La diferencia clave entre WCCN y NAP es la forma en que pesan las dimensiones en el espacio de los supervectores [52]. El método NAP elimina del todo algunas dimensiones a partir de proyectar los supervectores a un espacio de menor dimensión, mientras que WCCN más bien les da un peso determinado, lejos de eliminar completamente las dimensiones indeseadas.

3.2.4 i-vector

La modelación clásica empleando JFA basado en factores del canal y del locutor consiste en definir dos espacios diferentes: el del locutor definido por la matriz eigenvoice V y el espacio del canal definido por la matriz eigenchannel U.

Motivado recientemente por JFA una nueva técnica basada en la similaridad coseno entre vectores de baja dimensión ha sido introducida con el nombre de i-vector [30]. Esta aproximación evita la estimación por separado de la sesión y del locutor y es menos dependiente de la normalización de la puntuación, basándose en cambio, en las estadísticas de media y covarianza de un grupo de i-vectors de impostores.

La aproximación se basa en definir un único espacio de variabilidad, en lugar de dos separados. El nuevo espacio de Variabilidad Total contiene la variabilidad del canal y del locutor y es definido por la matriz T que contiene a los vectores propios de mayores autovalores de la matriz de covarianza [53]. Esta aproximación la motivan los experimentos realizados en [30] que constatan que los factores del canal del JFA los cuales supuestamente solo contenían los efectos del canal, eran portadores también de información del locutor.

En esta nueva aproximación se emplea JFA como un extractor de rasgos, que definirá el espacio T donde se compensará la varibilidad de canal, y que es de menor dimensión que el espacio del supervector de las GMM empleado en el clásico JFA.

Dado un locutor, el nuevo supervector GMM sería:

$$M = m + Tw, \tag{16}$$

[6] Todos sus autovalores son no negativos y los determinantes de sus submatrices principales también

donde m es el supervector independiente del canal y del locutor (que convenientemente es tomado como el vector de medias del UBM), T es la matriz rectangular de variabilidad total y w es un vector aleatorio con distribución normal estándar y cuyas componentes son los factores de variabilidad total.

Lo anterior implica que M está distribuído normalmente con media m y matriz de covarianza TT'. T se entrena de igual forma que la matriz de eigenvoice V, excepto por la importante diferencia de que para V todas las grabaciones de un mismo locutor se consideran como pertenecientes a él, sin embargo en el caso de T un conjunto de grabaciones de un único locutor son asumidas como producidas por distintos locutores.

El factor w es una variable oculta que es definida por su distribución a posteriori determinada por las estadísticas de Baum-Welch (BW). Esta distribución a posteriori es una gaussiana y su media es nuestro i-vector. Supongamos una secuencia de L ventanas $\{y_1, y_2, ..., y_L\}$ y un UBM Ω de C mezclas definidas en un espacio de rasgos de dimensión F, las estadísticas de BW necesarias para extraer el i-vector son obtenidas del UBM y son las siguientes:

$$N_c = \sum_{t=1}^{L} P(c|y_t, \Omega), \tag{17}$$

$$F_c = \sum_{t=1}^{L} P(c|y_t, \Omega) y_t, \tag{18}$$

donde $c = 1, ..., C$ es el índice de las Gaussianas y $P(c|x_t, \Omega)$ corresponde a la probabilidad posterior de la mezcla c generando el vector y_t. Con el objetivo de obtener el i-vector también es necesario calcular las estadística de BW de primer orden centralizada:

$$\tilde{F}_c = \sum_{t=1}^{L} P(c|y_t, \Omega)(y_t - m_c), \tag{19}$$

donde m_c es la media de la componente c del UBM. El i-vector para una secuencia u vendría dado por:

$$w = (I + T'\Sigma^{-1} N(u) T)^{-1} \cdot T'\Sigma^{-1} \tilde{F}(u), \tag{20}$$

donde $N(u)$ es una matriz diagonal de $CF \times CF$ cuyos bloques diagonales son $N_c I$ (c = 1,...,C). $\tilde{F}(u)$ es un supervector de dimensiones $C\dot{F}_1$ producto de concatenar las estadísticas de BW de primer orden $\tilde{F}_c$ para una secuencia dada u. Σ es una matriz de covarianza diagonal estimada durante el entrenamiento de FA, que modela la variabilidad residual que no fue capturada por la matriz de variabilidad total T.

4 Técnicas de normalización a nivel de puntuación

Afinar el umbral de decisión es una tarea bien complicada, debido principalmente a la variabilidad de los resultados de puntuación. Alguna de las causas de esta variabilidad son la naturaleza de los rasgos involucrados en el experimento, la cual puede variar entre los locutores, así como el contexto fonético, la duración, el ruido ambiental, calidad del modelo de entrenamiento del locutor. Tambien influye la diferencia entre los datos involucrados en la creación del modelo del locutor y los datos usados para la prueba. Lo cual es la problemática más desafiante en el reconocimiento del locutor [47].

La normalización de las puntuaciones va dirigida a aspectos tales como el escalado de las mismas y la normalización del auricular. El escalado de la distribución de las puntuaciones de diferentes locutores es usado para encontrar un umbral global independiente del locutor para el proceso de toma de decisión [28].

4.1 Modelo universal y de cohorte

Cuando los modelos estadísticos comenzaron a utilizarse para construir representaciones de los locutores, se observó que los valores de verosimilitud no eran confiables unas vez llegado el punto de decisión. Surgió la idea de que la regla de decisión dependiera no solo de los modelos de los locutores, si no también de modelos del conjunto de locutores fuera de los esperados, y comenzó a referirse como normalización de la puntuación [28]. Básicamente emergieron dos aproximaciones para modelar los locutores fuera del set: el modelo universal y el modelo de cohorte.

Asumamos que tenemos N locutores y sus correspondientes modelos estadísticos $\lambda_1, \lambda_2, ..., \lambda_N$. Si X denota al conjunto de vectores extraídos a una locución, de manera general la regla de decisión para un sistema cerrado de identificación de locutor pudiera expresarse:

$$P(X|\lambda^{ML}) \gtrless P(X|\lambda^{D}) \rightarrow X \in \begin{cases} \lambda^{ML} \\ \lambda^{D} \end{cases}, \tag{21}$$

donde $\lambda^{ML} = \lambda_i = argmax\{P(X|\lambda_n)\}$, y λ^{D} representa los modelos de los locutores desconocidos. De aplicar Bayes obtenemos la desigualdad:

$$\frac{P(X|\lambda^{ML})}{P(X|\lambda^{D})} \gtrless \frac{P(\lambda^{D})}{P(\lambda^{ML})} \rightarrow X \in \begin{cases} \lambda^{ML} \\ \lambda^{D} \end{cases}, \tag{22}$$

donde $\frac{P(S|\lambda^{ML})}{P(S|\lambda^{D})}$ es la puntuación que será calculada y $\frac{P(\lambda^{D})}{P(\lambda^{ML})}$ es el umbral que establecemos a priori.

La primera adaptación consiste en aproximar $P(X|\lambda^{D})$ con $P(X|\lambda^{UBM})$, donde λ^{UBM} es un modelo generado usando locuciones de una población grande de locutores, dicho modelo es conocido como Modelo Universal de Background (UBM). El modelo del locutor se obtiene mediante la actualización de los parámetros entrenados en el UBM [29]. Esto proporciona una conexión estrecha entre el modelo del locutor y el UBM llevando a un mejor rendimiento.

Alternativamente, en la normalización de cohorte, el modelo generado para cada locutor entrenado es asociado a un conjunto de locutores. El cohorte específico para cada locutor está formado por los N locutores de cohorte más similares a él, del conjunto de locutores que forman el cohorte.

En la actualidad, los sistemas automáticos de reconocimiento del locutor utilizan un conjunto de locutores impostores para el UBM y para la normalización que serían los locutores cohorte, con el objetivo de mejorar la robustez y la capacidad discriminativa del sistema. La idea es que los locutores impostores sean tomados como ejemplos negativos para el entrenamiento de un modelo discriminativo o utilizados para adaptar de un UBM el locutort. Durante la fase de verificación, la secuencia desconocida es comparada o calculada una puntuación repecto al UBM y respecto a los modelos de los locutores entrenados.

Por otro lado esta aproximación, empleando los conjuntos de cohorte, puede ser vista como un escalado para las distribuciones.

El uso de los modelos de background y de cohorte como una normalización a la verosimilitud o a las puntuaciones obtenidas a partir de las GMM, mejora sustancialmente los resultados.

4.2 Znorm

Znorm es uno de los métodos de normalización de la puntuación, propuesto por primera vez en [23]. Se usó masivamente en reconocimiento de locutores a mediado de los 90 [65]. En su propuesta, las variaciones de una determinada locución pueden ser eliminadas calculando la probabilidad logarítmica relativa a la media y varianza de la distribución de las puntuaciones de los impostores.

Esta técnica normaliza los modelos de los locutores teniendo en cuenta las diferentes condiciones de entrenamiento bajo los que fueron creados antes de la prueba. Znorm tiene dos formas, la primera es en la que cada modelo del locutor es asociada con un grupo de distribuciones de puntuación de impostores. Los parámetros (media y varianza) se calculan a partir de diferentes grupos de oraciones dichas por el impostor.

El modelo del locutor se prueba contra un grupo de señales de habla producidas por un impostor, obteniéndose una distribución de puntuación del impostor. La media y la varianza se estiman a partir de esta distribución y se aplica en puntuaciones similares obtenidas en un sistema de verificación. Una ventaja es que el estimado de los parámetros de normalizacion puede ser realizado previo a la prueba, o sea mientras el modelo del locutor se entrena.

Sea $L(x_i|S)$, la probabilidad logarítmica o puntuación para la i-ésima trama de la señal del locutor cuyo modelo es S, y el rasgo x_i, donde la señal completa la denotaremos por $X = \{x_i\}$, $i \in [1,N]$. De acuerdo a la notación anterior tenemos la siguiente ecuación:

$$L(X|S) = \frac{1}{N}\sum_{i=1}^{N} L(x_i|S), \tag{23}$$

y la puntuación normalizada

$$L_{norm}(X|S) = \frac{L(X|S) - \mu_I}{\sigma_I}, \tag{24}$$

donde I es el número de impostores y

$$\mu_I = \frac{1}{I}\sum_{i} LLR(x_i|S), \tag{25}$$

$$\sigma_I = \sqrt{\frac{1}{I}\sum_{i}(LLR(x_i|S) - \mu_I)^2}. \tag{26}$$

Znorm entonces calcula la puntuación del modelo S enfrentándolo a un conjunto de señales de impostores, μ_S^Z y σ_S^Z son la media y la desviación estándar derivadas del modelo del impostor S_i.

Znorm está relacionada con las diferencias en las condiciones de entrenamiento, de hecho busca alinear los modelos de los locutores generados bajo diversas condiciones de entrenamiento antes de la fase de prueba.

4.3 TNorm

Tnorm es otro método de normalización también basado en una estimación de la media y la varianza para el escalado. Durante la prueba un conjunto de modelos de impostores previamente obtenidos son usados para la obtención de la puntuación. Esta técnica ha sido muy utilizada en el Reconocimiento de Locutor independiente de texto y, aunque en menor medida, también en el dependiente de texto [28], [55]. La ventaja de Tnorm sobre una normalización de cohorte es el uso de la varianza, quien hace la aproximación de la distribución de los locutores de cohorte más exacta. La estimación de estos parámetros es llevada a cabo sobre los mismos segmentos de locución que con los que se hace la prueba, por lo tanto la diferencia acústica entre los segmentos de prueba y los normalizados, posibles con Znorm, son evitados con Tnorm. Otra significativa ventaja que tiene Tnorm sobre Znorm es la posibilidad que brinda comenzar la normalización previo a que se termine el entrenamiento. La normalización Tnorm viene dada por:

$$L_{norm}(x|S) = \frac{L(x|S) - \mu_{test}}{\sigma_{test}}, \tag{27}$$

donde μ_{test} y σ_{test} son la media y la desviación estándar de la distribución de puntuación de los impostores en el conjunto de rasgos de prueba.

Comparando Tnorm y Znorm se observa que estos métodos son muy similares y su diferencia radica en la manera de estimar los parámetros μ y σ.

Znorm solo precisa para estimar sus parámetros los modelos de los locutores obtenidos en el entrenamiento y los rasgos de los impostores, o sea que no hay que esperar a realizar la prueba para calcularlos. Esta es una ventaja que tiene sobre Tnorm, que además de necesitar los modelos de los impostores (que pueden obtenerse previamente) utiliza los rasgos de la señal de prueba y esto hace que no pueda aplicarse fuera de la estapa de prueba.

Sin embargo esta característica, coloca a Tnorm por encima de Znorm en muchas aplicaciones. Sucede que Tnorm lleva a cabo la estimación sobre los rasgos de prueba, por lo que el problema de la diferencia acústica entre la prueba y el entrenamiento presente en Znorm, aquí es evitado.

Znorm es generalmente considerada para compensar la variabilidad inter-locutor, mientras que Tnorm compensa la variabilidad de sesión. Como combinación de ambas técnicas es ampliamente usada ZT-norm, que no es más que Znorm seguida de Tnorm, y se emplea para compensar ambos. En cuanto a los requerimientos de memoria de cada técnica hay que tener en cuenta que mientras Znorm requiere del almacenamiento de los rasgos de los impostores, Tnorm precisa almacenar los modelos de los mismos.

4.4 HNorm

Este método, introducido en [19], se usa para lidiar con la desigualdad entre el auricular utilizado en el entrenamiento y el utilizado en la prueba. En [35] se emplea un identificador de auricular durante la prueba y de acuerdo a la clasificación que realice con cada señal a reconocer invoca al correspondiente conjunto de impostores para la normalización. Se infiere entonces que son necesarios conjuntos de impostores obtenidos utilizando diversos auriculares, lo que constituye su principal inconveniente.

Este método más allá de ayudar a normalizar respecto a la dependencia del auricular, permite fijar umbrales independientes del locutor más efectivos.

5 Configuraciones

Actualmente la concepción de los sistemas de reconocimiento de locutor ha sido integrar muchos de sus métodos y herramientas, haciendo bastante difuso el radio de acción de cada una. En la actualidad, es muy difícil diferenciar el concepto de rasgos del de modelos, cuando se utilizan los supervectores de media de los clasificadores GMM como rasgos introducidos a las máquinas de soporte vectorial.

De igual forma, en el caso del análisis de factores, clasificado en este trabajo como un método de normalización a nivel de modelo, hay sistemas muy modernos que explotan su utilización antes de la clasificación, y es visto este como un extractor de rasgos [54]. La búsqueda de Kernels para las SVM ha llevado a la introducción del kernel coseno para el cálculo de la disimilaridad de los i-vectors y un salto cualitativamente importante introducido en [30] ha llevado a muy buenos resultados en la clasificación empleando el cálculo de la distancia coseno entre los modelos entrenados y los locutores de prueba, directamente como una puntuación de decisión, lo que eliminaría a las SVM del proceso de decisión y se evitaría el entrenamiento de las mismas haciendo el proceso más rápido y menos complejo.

Desde que el UBM forma parte de la mayoría de los sistemas de reconocimiento de locutores, supone una forma natural de crear supervectores. Esto ha llevado a la conformación de clasificadores híbridos

donde los modelos generativos GMM-UBM son empleados para crear vectores de las discriminativas SVMs.

Resumiendo, dos aspectos esenciales para el diseño de un reconocedor basado en supervectores son (1) cómo crear el SV a partir de la secuencia de entrada, (2) cómo estimar y aplicar la compensación de varibilidad de sesión en el dominio del SV. Sin perder de vista cómo se realizará el cálculo de los scores una vez compensados los modelos.

5.1 Supervectores GLDS para SVM

Uno de los métodos más simples para obtener supervectores es GLDS (Generalized Linear Discriminant Sequence) [58]. Este método crea los supervectores mapeando directamente al espacio del kernel utilizando una expansión polinomial [7]. Durante el entrenamiento, los locutores impostores y los de entrenamiento $X = \{x_1, x_2, ..., x_T\}$ son representados como el promedio de vectores de rasgos expandidos:

$$b_{prom} = \frac{1}{T}\sum_{t=1}^{T} b(x_t). \tag{28}$$

Los vectores promedio están normalizados respecto a la varianza utilizando los locutores del modelo universal, y fueron designados con la etiqueta apropiada para el entrenamiento SVM ($+1$ =locutores target; -1 =locutores del modelo universal). El mayor inconveniente de GLDS es lo complicado que se vuelve controlar la dimensionalidad de los supervectores, en la práctica la expansión polinomial incluye monomios de 2do y 3er orden antes de que el problema de la dimensionalidad sea inviable [47].

5.2 Kernel gaussiano para SVM (*GSV-SVM)*

Como dijimos anteriormente en el estado del arte el modelo generativo GMM-UBM es muy usado en la creación de los supervectores que luego serán clasificados con las SVMs. En [38] los autores obtienen el kernel supervector gaussiano delimitando la distancia de Kullback-Leibler (KL) entre las gaussianas. Supongamos que tenemos un UBM, $\lambda_{UBM} = \{P_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ y dos secuencias a y b que vienen representadas por sus GMM MAP-adaptadas, $\lambda_a = \{P_k, \mu_k^a, \Sigma_k\}_{k=1}^{K}$ y $\lambda_b = \{P_k, \mu_k^b, \Sigma_k\}_{k=1}^{K}$ (nótese como su única diferencia radica en las medias). Luego el kernel de la divergencia KL es definido:

$$K(\lambda_a, \lambda_b) = \sum_{k=1}^{K} (\sqrt{P_k}\Sigma_k^{(-1/2)}\mu_k^a)'(\sqrt{P_k}\Sigma_k^{(-1/2)}\mu_k^b). \tag{29}$$

Desde el punto de vista de implementación, esto significa que todas las medias gaussianas μ_k tienen que ser normalizadas con $\sqrt{P_k}\Sigma_k^{(-1/2)}$ antes de que pasen a formar parte del entrenamiento de la SVM, esto es una forma de normalización de la varianza. Por lo tanto y a pesar de eso, solo los vectores de media de la GMM están incluídos en el SV, la información contenida en la varianza y peso de las GMM está implícitamente presente en la normalización del supervector gaussiano.

5.3 Supervector MLLR

En [59], [52] los autores emplean los parámetros del método de regresión lineal de máxima probabilidad (MLLR de sus siglas en inglés) como entrada a las SVMs. MLLR transforma los vectores de media de un

[7] Una expansión polinomial de 2do orden para un vector de 2 dimensiones $x = (x_1, x_2)'$ viene dada por $b(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)'$

modelo independiente del locutor como $\mu'_k = A\mu_k + b$, donde μ'_k es el vector de media adaptado, μ_k es el vector de media del UBM y los parámetros A y b definen la transformación lineal. Estos parámetros son estimados maximizando la probabilidad de los datos de entrenamiento con un algoritmo de maximización de la expectancia [8] (EM) modificado.

5.4 ¿Qué supervector usar entonces?

Dadas las múltiples opciones para crear un supervector y modelar la variabilidad de sesión, ¿cuál elegir para las aplicaciones prácticas? Resulta complicado comparar los métodos en la literatura debido a las diferencias en la selección de los datos, el establecimiento de los parámetros y entre otros detalles de implemetación [47]. Sin embargo hay algunas prácticas comunes que pueden seguirse, para ello presentamos las evaluaciones NIST 2008 de reconocimiento del locutor realizadas por el consorcio I4U [60]. Todos los clasificadores utilizan rasgos espectrales short-term y el centro de la investigación fueron los clasificadores de supervectores. Tres conocidos métodos GMM-UBM [46], GLDS [58] y GSV-SVM [38] fueron estudiados. Los resutados se muestran en la Tabla 1. Además fueron propuestos nuevos kernels: transformación de rasgos (FT-SVM [63]), kernel de secuencia probabilística (PSK-SVM [64]) y el kernel Bhattacharyya (BK-SVM [61]).

Tabla 1. Desempeño de los clasificadores individuales y fusionados [60].

	EER (%)
Modelos de Mezclas Gaussianas (GMM)	
GMM-UBM	8.10
GMM-UBM (eigenchannel)[40]	5.22
GMM-UBM (JFA) [40]	3.11
SVM con diferentes kernels	
GLDS-SVM [58]	4.44
GSV-SVM [38]	4.43
BK-SVM [61]	2.05
Fusión de GMM-UBM(eigenchanel) con BK-SVM	2.05

El cuadro 1 muestra el desempeño de los sistemas individuales junto a la fusión con clasificadores. La exactitud es medida en función del EER (*equal error rate*), una medida del error de verificación que da la exactitud del umbral para el cual los falsos rechazos y las falsas aceptaciones son iguales [47]. De los resultados tabulados es evidente que la compensación de sesión mejora significativamente la exactitud del sistema GMM-UBM. También puede ser visto como, dentro de los clasificadores individuales, GMM-UBM con JFA es el mejor, y que JFA se desempeña mejor que el método eigenchannel (que es un caso particular de JFA). Finalmente fusionando los clasificadores con compensación de canal se mejora la exactitud como era esperado.

[8] Algoritmo que estima los parámetros del modelo del locutor maximizando la probabilidad de que la GMM modele la clase de dicho locutor

6 Experimentación

Como parte del estudio y familiarización con las técnicas para compensar la variabilidad de sesión, se realizaron algunos experimentos. Uno de los objetivos principales de la experimentación lo constituye la exploración del software ALIZE, en particular de las herramientas relacionadas con FA, sin embargo aunque los resultados no fueron los esperados se exponen en la siguiente sección. Se llevaron a cabo también un conjunto de experimentos alternativos que permitieron profundizar en el estudio de los métodos de normalización de la puntuación y de culminar con interesantes aportes en un trabajo presentado en RECPAT 2011 [57].

6.1 Experimentos con FA

Alize [9] es una herramienta para el reconocimiento automático del locutor, como tal contiene algoritmos que pueden identificar a personas por la voz. El paquete LIA-RAL que emplea Alize, fue la interfaz usada para la experimentación, ambos paquetes continúan en constante desarrollo. El empleo de esta herramienta nos permitió constatar que existe muy poca documentación acerca de cómo experimentar con esta técnica, no obstante Alize constituye probablemente el más completo y actualizado software libre con estos fines [47].

Buscando obtener, con las las funcionalidades del ALIZE, un sistema verificador de locutores que tenga incluido el análisis de factores como herramienta para enfrentar la variabilidad de sesión, se hizo necesario emplear una base de datos que tuviera locutores grabados en varias sesiones y canales. La base de datos utilizada fue NIST 2001 Ahumada [56], una base de 103 grabaciones de locutores masculinos, obtenidas bajo condiciones controladas para la caracterización e identificación.

Para la explotación de la aproximación FA implementada en ALIZE, se seleccionaron de Ahumada las sesiones a utilizar, y se distribuyeron adecuadamente en las tareas de entrenamiento y prueba. Los resultados fueron comparados con la línea base, que reproducía los pasos de obtención de los rasgos pero que no modelaba los locutores teniendo en cuenta FA. En las figuras 3 y 4 se muestran las matrices de verosimilitud de la línea base y del experimento que incorpora FA respectivamente. Estas matrices contrastan los modelos de los 50 locutores entrenados con esos mismos 50 locutores durante la prueba, y donde los tonos azules/rojos representan las puntuaciones más bajas/altas .

Como se observa, la aplicación de FA distorsiona los resultados respecto a la línea base asignándole valores de probabilidad altos a locutores fuera de la diagonal. Esto no era lo esperado y la más fuerte causa es la limitada cantidad de locutores para enriquecer la variabilidad del entrenamiento.

Igualmente como parte de los objetivos propuestos, se hacen experimentos con el algoritmo de JFA desarrollado por la Facultad de Tecnología Informática de la Universidad de Brno, Checoslovaquia, buscando comprender los procesos de estimación de los factores ocultos y la obtención de las matrices.

Durante un proceso de verficación usando GMM-UBM, hay un grupo de pasos que son comunes a la mayoría de los sistemas. Cada ventana es representada por vectores de rasgos de dimensión 24 ($MFCC + \Delta$), son extraídos los silencios y normalizados con la sustracción de la media y la varianza. El modelo fue entrenado con 512 mezclas, empleando el algoritmo de máxima expectancia y los modelos target fueron adaptados con MAP.

[9] Ahora bajo la plataforma biométrica de autentificación “Mistral”. Disponible en: http://mistral.univ-avignon.fr/en/.

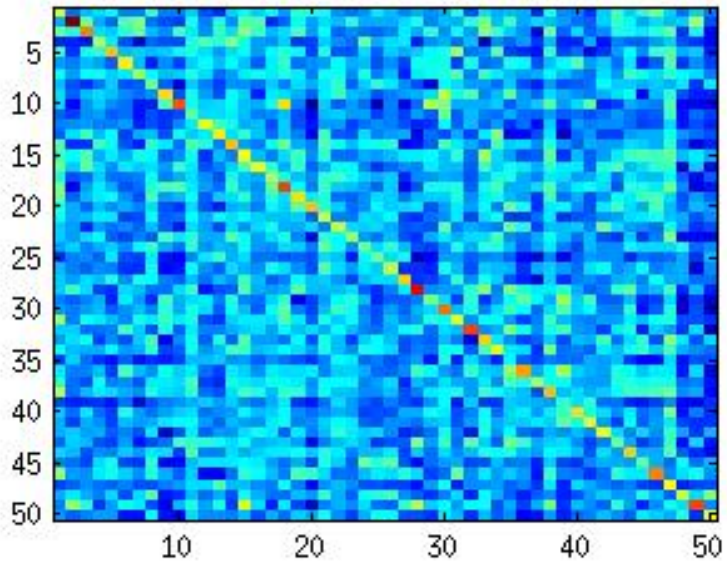

Fig. 3. Matriz de verosimilitud de la línea base.

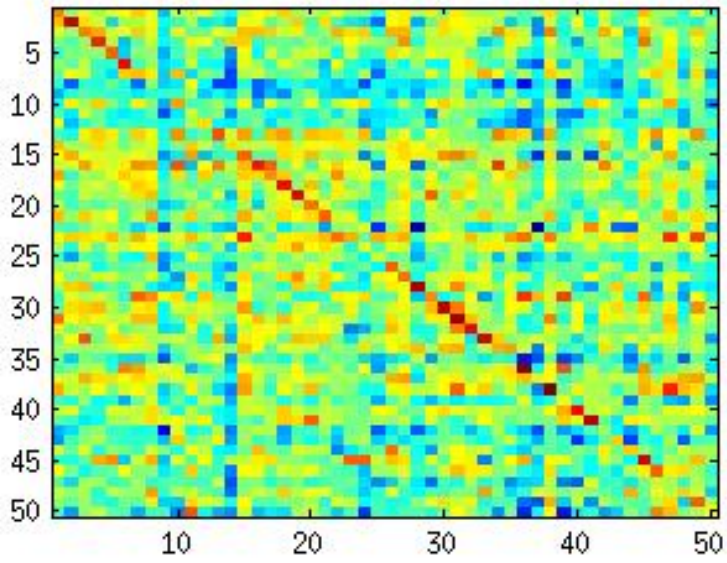

Fig. 4. Matriz de verosimilitud del experimento con FA.

6.2 Nuevo método para la normalización de la puntuación

Basados en el método de normalización de la puntuación Hnorm y en las significativas mejoras que implica eliminar las diferencias de canal entre la sesión de entrenamiento y prueba, proponemos un nuevo método de normalización que llamamos L_{norm} y que describiremos a continuación.

La idea del primer grupo de experimentos es que basados en características del canal y el auricular por el que fueron obtenidas las señales de prueba, se haga una selección del grupo de locutores impostores que se emplearán en la normalización de la puntuación, buscando cotejar estas características. Si bien es cierto que en la mayoría de las aplicaciones no se cuenta con información a priori del canal del que proviene la señal de prueba, los experimentos demostraron que el uso del conjunto de cohorte con mejores[10] valores de relación señal-ruido, atenuación de la señal y sensibilidad, elevan significativamente el rendimiento del sistema.

Tabla 2. Canal cotejado de las señales de prueba y de los impostores.

	Znorm		Tnorm		ZTnorm	
Características	DCF	EER	DCF	EER	DCF	EER
Linea base	2.45	8	3.49	12	3.20	6
Alta S/N	0.86	1.61	0.54	1.72	0.75	0.97
Baja atenuación	2.52	3.16	2.04	2.37	3.04	4.21
Alta sensibilidad	2.68	5.69	2.68	3.85	3.74	5.38

Producto de esta experimentación observamos cómo la mejora es muy significativa al hacer coincidir las características estudiadas del canal y el auricular entre entrenamiento y prueba, aún cuando estas no fueran buenas, si coincidían se obtenían mejores resultados, lo que significa que tiene mayor impacto en el sistema la variabilidad de sesión que la calidad de la señal.

Se observó también cómo Tnorm es mucho más sensible a las variaciones de las señales de prueba (lo que está en correspondencia con su definición) y esto es importante a la hora de seleccionar el método de normalización. Y finalmente de las tres características estudiadas resultó ser la relación señal-ruido la que más influyó en los resultados.

Buscando acercarnos más a las condiciones de situaciones reales, en las que no se tiene conocimiento de las características del canal por el que viaja la señal de prueba, se realizaron los mismos experimentos pero con señales de prueba con característcas genéricas del canal. Los resultados se muestran en 3.

Tabla 3. Señales de prueba con características genéricas del canal.

	Znorm		Tnorm	
Características	DCF	EER	DCF	EER
Linea base	2.45	8	3.49	12
Alta S/N	3.04	7.43	2.29	4.16
Baja atenuación	3.21	6	3.14	5.43
Alta sensibilidad	3.13	6.41	3.37	7.22

En el segundo grupo de experimentos realizados, la mejoría no fue tan marcada pero sí significativa. Esto prueba que es mejor normalizar utilizando grupos de impostores pequeños pero obtenidos bajo buenas condiciones, que emplear un grupo mayor sin seleccionar.

[10] Alta relación señal-ruido, alta sensibilidad y baja atenuación

7 Conclusiones

Durante los últimos 10 años, la comunidad de reconocimiento del locutor ha hecho significativos avances tecnológicos. En [47] se hace una selección de las técnicas más influyentes que han demostrado funcionar en la práctica. Nos referimos a los dirigidos a robustecer los sistemas en cuanto a la variabilidad de sesión, ya que la búsqueda de una compensación de sesión discriminativa es ciertamente una interesante dirección para futuros estudios. La incorporación de la adaptación a partir del UBM [46], los métodos de normalización, calibración y fusión de la puntuación [28] y la modelación explícita de la variabilidad de sesión y su compensación [50], [40] definen avances con probada efectividad y eficiencia en las competencias NIST. Aunque efectivos, el mayor inconveniente de estos métodos es su costo en cuanto a la masiva cantidad de data requerida para entrenar los modelos de background, de cohorte para la normalización del score y para la modelación de la variabilidad de sesión y de locutor. Estos datos deben ser etiquetados y organizados de manera controlada, requiriendo significativos esfuerzos humanos. Si las características de la data no se corresponden con los requerimientos necesarios para la operación , sucede que la efectividad del sistema cae marcadamente, incluso a niveles no utilizables. Para transferir esta tecnología a la práctica, es necesario dirigir esfuerzos a hacer los métodos menos dependientes de la selección del conjunto de datos. Los métodos también exigen simplificación computacional.

Finalmente, las técnicas actuales precisan mucho tiempo de entrenamiento y prueba para dar resultados satisfactorios, lo que es un gran reto para las aplicaciones que tengan que tomar una decisión en tiempo real.

Referencias bibliográficas

1. R. Mammone, X. Zhang, and R. Ramachandran, “Robust speaker recognition: a feature based approach”. IEEE Signal Processing Magazine 13, September 1996.
2. B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification”, J.Acoust. Soc. Amer., vol. 55, pp. 1304-1312, 1974.
3. H. Hermansky and N. Morgan, “RASTA processing of speech”, IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 578-589, Oct. 1994.
4. J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification”, in Proc. 2001: A Speaker Odyssey, pp. 213-218, 2001.
5. B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinat, “Short-time gaussianization for robust speaker verification”,in Proc. ICASSP’02 vol. 1, pp. 681-684, 2002.
6. Gray C.H. and Kopp G.A., “Voiceprint Identification”. Bell Telephone Laboratories Report, Bell Laboratories, 1944.
7. M. Mason, R. Vogt, B. Baker, and S. Sridharan, “Data-driven clustering for blind feature mapping in speaker verification”, in Proc. Interspeech 05, pp. 3109-3112, 2005.
8. O. Thyes, R. Kuhn, P. Nguyen, and J. C. Junqua, “Speaker identification and verification using eigenvoices”, in Proc. ICSLP’00 pp.242-245, 2000.
9. P. Kenny, M. Mihoubi, and P. Dumouchel, “New map estimators for speaker recognition”, in Proc. Eurospeech’03, pp. 2964-2967, 2003.
10. P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data”, IEEE Trans. Speech Audio Process., vol. 13,no. 3, pp. 345-354, May 2005.
11. R. Vogt, B. Baker, and S. Sridharan, “Modelling session variability in text-independent speaker verification”, in Proc. Interspeech’05, pp. 3117-3120, 2005.
12. A. Solomonoff, W. Campbell, and C. Quillen, “Channel compensation for SVM speaker recognition”, in Proc. Odyssey04, pp. 57.62, 2004.
13. A. Solomonoff, W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition”, in Proc. ICASSP’05, pp. I-629-I-632, 2005.
14. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Improvements in factor analysis based speaker verification”, in Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 113-116, Toulouse, May 2006.

15. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor analysis simplified”, in Proceedings of the IEEE Conference on Acoustics,Speech, and Signal Processing, vol. 1, pp. 637-640, Philadelphia, Mar 2005.
16. P. Kenny and P. Dumouchel, “Experiments in speaker verification using factor analysis likelihood ratios”, in Proc.Odyssey04, pp. 219.226, 2004.
17. A. O. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-bsaed speaker recognition”, in Proceedings of the International Conference on Spoken Language Processing, pp. 1471-1474, Pittsburgh, PA, Sep.2006.
18. Douglas A Reynolds, “Channel robust speaker verification by feature mapping”, Acoustics,Speech and Signal Processing 2003.
19. Reynolds, D.A. “The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus”, Proceedings of IEEE ICASSP ’96, Vol. 1, pp. 113-116, 1996.
20. Wu D., Li, J. and Wu H., “Improving text-independent speaker recognition with locally nonlinear transformation”, Technical report, Computer Science and Engineering Department, York University, Canada, 2008.
21. Furui Sadaoki, “Cepstral analysis technique for automatic speaker verification”.IEEE Transactions on speech and audio processing,Vol.ASSP-29,No.2 .April 1981.
22. Liu F., Stern R., Huang X. and Acero A. “Efficient Cepstral Normalization for Robust Speech Recognition”. Proceedings of ARPA Human Language Technology Workshop, March 1993.
23. Li, K. P. and Porter, J. E., “Normalizations and selection of speech segments for speaker recognition scoring”. Proceedings of ICASSP ’88, Vol. 1, pp. 595-598, 1988.
24. Rosenberg, A.E. and Parthasarathy, S., “Speaker background models for connected digit password speaker verification”. Proceedings of ICASSP ’96, Vol. 1, pp. 81-84, 1996.
25. Jin Q. y Waibel A., “Application of LDA to Speaker Recognition”. Proc. of the ICSLP, Beijing, China, October 2000.
26. Sun, Z. P., Mason J. S., “Combining features via LDA in speaker recognition”, 1993.
27. Wu D., Li J. and Jiang H., “Normalization and Transformation Techniques for Robust Speaker Recognition”, Speech Recognition, Technologies and Applications, I-Tech, Vienna, Austria, ISBN 978-953-7619-29-9, pp. 550, November 2008.
28. Auckenthaler R., Carey M., and Lloyd-Thomas H., “Score normalization for text independent speaker verification system”, Digital Signal Processing, vol. 10, pp. 42-54, Jan 2000.
29. Fortuna J., Sivakumaran P., Ariyaeeinia M. and Malegaonkar A., “Relative effectiveness of score normalisation in open-set speaker identification”, University of Hertfordshire, ODYSSEY04 The Speaker and Language Recognition Workshop Toledo, Spain May 31-June 3, 2004.
30. Najim Dehak, “Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification”, Ph.D. thesis, Ecole de Technologie Superieure, Montreal, 2009.
31. A. de la Torre, J. C. Segura, M. C. Benitez, A. M. Peinado and A. J. Rubio, “Non-linear transformations of the feature space for robust speech recognition”. In Proceedings of ICASSP02, Orlando, pp 401-404, 2002.
32. C. Hsu and L. Lee, “Extension and further analysis of higherorder cepstral moment normalization for robust features in speech recognition”. INTERSPEECH 2006.
33. D. A. Reynolds, “An overview of automatic speaker recognition technology”. In proceedings of ICASPSP‘02, 2002.
34. S. Dharanipragada and M. Padmanabhan, “A nonlinear unsupervised adaptation technique for speech recognition”, In ICSLP’00, vol.4, 556-559, 2000.
35. D. A. Reynolds, “Comparison of background normalization methods for text-independent speaker verification”, Proceedings of Eurospeech, 1997.
36. Markel, J., Oshika, B., and A.H. Gray, j. “Long-term feature averaging for speaker recognition”. IEEE Trans. Acoustics, Speech, and Signal Processing 25, 330-337, August 1977.
37. Kinnunen, T., Hautam Ì ki, V., and Franti, P. “On the use of longterm average spectrum in automatic speaker recognition”. In 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLPâ06) , pp. 559-567, Singapore, December 2006.
38. Campbell, W., Sturim, D., and Reynolds, D. “Support vector machines using GMM supervectors for speaker verification”. IEEE Signal Processing Letters 13, 5, 308-311, May 2006.
39. Burget, L., Matjka, P., Schwarz, P., Glembek, O., and Cernock, J. Ì“Analysis of feature extraction and channel compensation in a GMM speaker recognition system”. IEEE Trans. Audio, Speech and Language Processing 15, 7, 1979-1986, September 2007.
40. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. “A study of inter-speaker variability in speaker verification”. IEEE Trans. Audio, Speech and Language Processing 16, 5, 980-988, July 2008.
41. Vogt, R., and Sridharan, S. “Explicit modeling of session variability for speaker verification”. Computer Speech and Language 22, 1, 17-38, January 2008.
42. Kenny, P., Boufianne, G., Ouellet, P., et Dumouchel, P. “Joint Factor Analysis versus Eigenchannels in Speaker Recognition”. IEEE Transaction on Audio Speech and Language Processing, 15(4): 1435-1447, 2007.
43. Kenny, P., Boulianne, G., Ouellet, P., et Dumouchel, P. “Speaker and Session Variability in GMM-Based Speaker Verification”. IEEE Transaction on Audio Speech and Language Processing, 15(4): 1435-1447, 2007.

44. N. Dehak, P. Kenny, and P. Dumouchel, “Modeling prosodic features with joint factor analysis for speaker verification”. IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 7, pp. 2095-2103, Sept. 2007.
45. N. Dehak, P. Kenny, and P. Dumouchel, “Continuous prosodic features and formant modeling with joint factor analysis for speaker verification”. In Proceedings of INTERSPEECH’07. pp.1234-1237, 2007.
46. Reynolds, D., Quatieri, T., and Dunn, R. “Speaker verification using adapted gaussian mixture models”. Digital Signal Processing 10, 1, 19-41, January 2000.
47. Tomi Kinnunen, Haizhou Lib, “An overview of text-independent speaker recognition: from features to supervectors”, Speech Communication 52, 12-40, 2010.
48. Remco Teunen, Ben Shahshahani, and Larry Heck, “A model-based transformational approach to robust speaker recogntion”, ICSLP, 2000.
49. Wei Wu, Thomas Fang Zheng, and Mingxing Xu, “Cohort-based speaker model synthesis for channel robust speaker recognition”, ICASSP 2006.
50. Brummer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., Leeuwen, D., Matejka, P., Schwartz, P., Strasheim, A. “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006”. IEEE Trans. Audio, Speech and Language Processing 15, 7, 2072-2084, September 2007.
51. Fauve, B., Matrouf, D., Scheffer, N., Bonastre, J.-F., and Mason, J. “State-of-the-art performance in text-independent speaker verification through open-source software”. IEEE Trans. Audio, Speech and Language Processing 15, 7, 1960-1968, September 2007.
52. Stolcke, A., Kajarekar, S., Ferrer, L., and Shriberg, E. “Speaker recognition with session variability normalization based on MLLR adaptation transforms”. IEEE Trans. Audio,Speech and Language Processing 15, 7, September 2007.
53. N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification”, in Interspeech, Brigthon, 2009.
54. N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, “Front-end Factor Analysis for Speaker Verification”, submitted to IEEE Transaction on Audio, Speech and Language Processing, 2009.
55. Navratil, J. and Ramaswamy, G. N., “The awe and mystery of t-norm”, Proceedings of Eurospeech, 2009-2012, 2003.
56. Ortega, J., Gonzalez, J. and Marrero, V., “AHUMADA: A large speech corpus in Spanish for speaker characterization and identi- fication”, Speech Communication, vol 31, 255-264, 2000.
57. Montalvo, A., Calvo, J., “Influence of Channel and Handset Characteristics in the Normalization of the Speaker Verification Score”, RECPAT 2011.
58. Campbell, W., Campbell, J., Reynolds, D., Singer, E., and Torres-Carrasquillo, P. “Support vector machines for speaker and language recognition”. Computer Speech and Language 20, 2-3, 210-229, April 2006.
59. Karam, Z., and Campbell, W. “A new kernel for SVM MLLR based speaker recognition”. In Proc. Interspeech’07 (ICSLP), pp. 290-293, Antwerp, Belgium, August 2007.
60. Li, H., Ma, B., Lee, K.-A., Sun, H., Zhu, D., Sim, K., You, C., Tong, R., Karkkainen, I., Huang, C.-L., Pervouchine, V., Guo, W., Li, Y., Dai, L., Nosratighods, M., Tharmarajah, T., Epps, J., Ambikairajah, E., Chng, E.-S., Schultz, T., and Jin, Q. “The I4U system in NIST 2008 speaker recognition evaluation”. In Proc. Int. conference on acoustics, speech, and signal processing ICASSP’09, pp. 4201-4204, Taipei, Taiwan, April 2009.
61. You, C., Lee, K., and Li, H. “An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition”. IEEE Signal Processing Letters 16, 1, 49-52, January 2009.
62. Burget, L., Brummer, N., Reynolds, D., Kenny, P., Pelecanos, J., Vogt, R., Castaldo, F., Dehak, N., Dehak, R., Glembek, O., Karam, Z., Noecker, J., Na, E., Costin, C., Hubeika, V., Kajarekar, S., Scheffer, N., and Cernock J., “Robust speaker recognition over varying channels” Report from JHU workshop 2008. Technical report, March 2009.
63. Zhu, D., Ma, B., and Li, H. “Joint MAP adaptation of feature transformation and gaussian mixture model for speaker recognition”. In Proc. Int. conference on acoustics, speech, and signal processing, pp. 4045-4048, ICASSP’09, Taipei, Taiwan, April 2009.
64. Lee, K., You, C., Li, H., Kinnunen, T., and Zhu, D. “Characterizing speech utterances for speaker verification with sequence kernel SVM”. In Proc. 9th Interspeech’08, pp. 1397-1400 Brisbane, Australia, September 2008.
65. Bimbot F., Bonastre J., Fredouille C., Gravier G., Meignier S., Merlin T., Ortega-Garc ÌJ., Magrin I., Petrovska D. and Reynolds D. “A Tutorial on Text-Independent Speaker Verification”. EURASIP Journal on Applied Signal Processing 2004:4, 430-451, 2004.
66. Burget, L., Brummer, N., Reynolds, D., Kenny, P., Pelecanos, J., Vogt, R., Castaldo, F., Dehak, N., Dehak, R., Glembek, O., Karam, Z., Noecker, J., Na, E., Costin, C., Hubeika, V., Kajarekar, S., Scheffer, N., and Cernock J., “Robust speaker recognition over varying channels” Report from JHU workshop 2008. Technical report, March 2009.

RT_051, noviembre 2012
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
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;
La Habana. Cuba. C.P. 12200
Impreso en Cuba