Verificación e Identificación de Locutor con Normalización de Vectores de Características en Entornos Acústicos Adversos

Luis Buera[1], Eduardo Lleida[1], Juan Diego Rosas[1], Jesús Villalba[1], Antonio Miguel[1], Alfonso Ortega[1], Óscar Saz[1]

[1] Departamento de Electrónica y Comunicaciones. Instituto de Investigación en Ingeniería de Aragón, I3A. Universidad de Zaragoza
{lbuera, lleida, jdrosas, villalba, amiguel, ortega, oskarsaz}@unizar.es

Abstract. En condiciones acústicas adversas, el comportamiento de los sistemas de verificación e identificación de locutor se degrada significativamente. Para compensar este hecho, se suelen aplicar distintas técnicas, una de las cuales se presenta en este trabajo. El algoritmo "Phoneme Dependent Multi-Environment Models based Linear Normalization", PD-MEMLIN, se utiliza en una fase previa a la verificación o identificación de locutor para compensar así los efectos negativos del ruido. En esta técnica los vectores de características se modelan mediante una mezcla de Gaussianas, GMM ("Gaussian Mixture Models") para cada fonema de los espacios limpio y ruidoso. Por otra parte, PD-MEMLIN estima una transformación lineal asociada a cada par de Gaussianas del mismo fonema, entendiendo por cada par una Gaussiana de la GMM del espacio de vectores de características limpios y otra de la GMM del espacio de los ruidosos. Para comprobar la mejora de esta técnica en tareas de verificación e identificación de locutor se llevaron a cabo distintos experimentos con la base de datos SpeechDat Car: utilizando un sistema de verificación de locutor UBM-GMM se obtuvo una mejora en términos de EER, "Equal Error Rate", del 70.2%, mientras que empleando un sistema de identificación de locutor basado en GMM la mejora alcanzó el 48.69%.

Keywords: Normalización de vectores de características, verificación e identificación de locutor bajo condiciones acústicas adversas, GMM.

1 Introducción

En condiciones acústicas adversas, el comportamiento de los sistemas de verificación e identificación de locutor se degrada significativamente. Para compensar este hecho se suelen aplicar distintas técnicas [1]. Simplificando, en aquellos sistemas de verificación e identificación de locutor basados en GMM, "Gaussian Mixture Models", se pueden considerar dos tipos de técnicas para proporcionar robustez, a saber, la adaptación de los modelos y la normalización de vectores de características.

El primero de los tipos, que sólo modifica los parámetros de las distintas Gaussianas que componen los modelos estadísticos, puede ser más específico, mientras que las técnicas de normalización de vectores de características requieren de menos datos y tiempo de computación. Por otra parte, en muchas ocasiones ambos tipos de técnicas de robustez pueden emplearse conjuntamente, en cuyo caso se procedería a compensar inicialmente los vectores de características para posteriormente adaptar los modelos con la señal ya compensada. Sin embargo en entornos acústicos reales, muchas veces extremadamente variantes, es imposible reentrenar los modelos para cada situación acústica concreta, de modo que en esos casos la normalización de los vectores de características se convierte en la única opción válida para dotar al sistema de la robustez requerida.

Las técnicas de compensación de vectores de características se pueden dividir en tres grandes grupos [2]: compensación basada en modelos, compensación empírica y compensación mediante filtro paso alto en el dominio cepstral. El primer grupo emplea un modelo matemático para representar el efecto del ruido, cuyos parámetros deberán estimarse mediante las tramas degradadas. Técnicas como "Vector Taylor Series" para normalización, VTS, [3], o "Codeword Dependent Cepstral Normalization", CDCN, [4], son ejemplos de este tipo de compensaciones basadas en modelos predeterminados. Por su parte, la compensación empírica no asume ningún tipo de modelo de degradación y emplea señal limpia y contaminada, generalmente estéreo aunque no es imprescindible, para determinar la correspondiente degradación. Ejemplos de este grupo son "multivariate gaussian-based cepstral normalization", RATZ, [3], "Stereo based Piecewise LInear Compensation for Environments", SPLICE, [5], o "Multi-Environment Models based LInear Normalization", MEMLIN, [6]. Por último los métodos basados en compensación mediante filtro paso alto en el dominio cepstral no obtienen resultados tan satisfactorios como los logrados con las técnicas anteriormente comentadas, pero en muchas ocasiones son utilizados por tener un coste computacional casi nulo. El algoritmo más empleado de este último grupo es "Cepstral Mean Normalization", CMN, [2].

En este trabajo se va a emplear una técnica de compensación empírica ("Phoneme Dependent Multi-Environment Models based LInear Normalization", PD-MEMLIN) para mejorar el comportamiento de los sistemas de verificación e identificación de locutor bajo condiciones acústicas adversas. Este algoritmo hace uso del estimador MMSE, "Minimum Mean Square Error", utilizando a su vez una serie de transformaciones lineales aprendidas en una fase de entrenamiento previa. Anteriormente a ello, el espacio formado por los vectores de características limpios se modela con una GMM para cada fonema, del mismo modo que el espacio formado por los vectores de características ruidosos. De este modo, cada una de las transformaciones lineales aprendidas estará asociada a un par de Gaussianas del mismo fonema, lográndose así una técnica de normalización de los vectores de características capaz de compensar los efectos de un entorno acústico adverso y variable.

Este trabajo se organiza del siguiente modo: en la Sección 2, se presenta la técnica PD-MEMLIN. Los sistemas de verificación e identificación de locutores empleados

se explican en la Sección 3. Los experimentos y resultados se incluyen en la Sección 4. Finalmente se comentan las conclusiones en la Sección 5.

2 Phoneme Dependent MEMLIN

"Phoneme Dependent Multi-Environment Models based LInear Normalization", PD-MEMLIN, es una técnica de normalización de vectores de características empírica que requiere de señal estéreo a la hora de determinar las distintas transformaciones lineales propuestas. El espacio formado por los vectores de características limpios se modela mediante una GMM para cada fonema, al igual que el espacio formado por los vectores de características ruidosos, que además se divide en varios entornos básicos. Las transformaciones lineales se estiman entre cada par de Gaussianas del mismo fonema, entendiendo por cada par, una Gaussiana del espacio de la señal limpia y otra del espacio de la señal ruidosa, siempre del mismo fonema. Puede apreciarse un esquema para un solo entorno en la Figura 1.

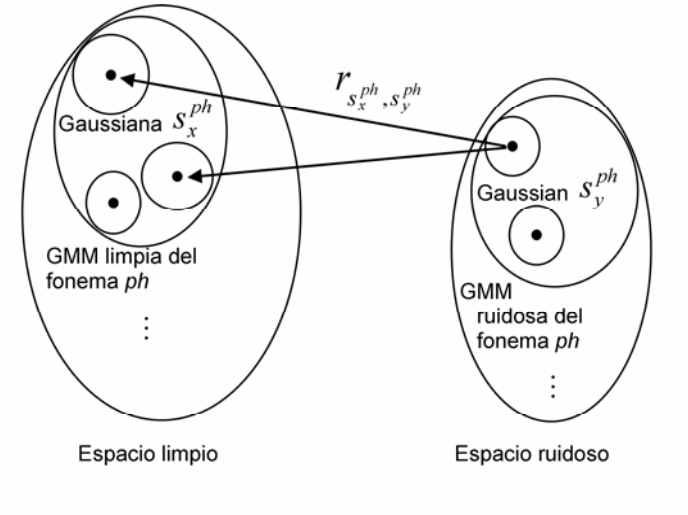

Fig. 1. Visión gráfica de PD-MEMLIN para un único entorno básico.

2.1 Estimador MMSE

Dado un vector de características ruidoso para cada instante de tiempo, t, y_t, el correspondiente vector de características limpio estimado, $\hat{x}_t$, puede calcularse mediante el estimador MMSE, donde x es el vector de características limpio

$$\hat{x}_t = E\left[x|y_t\right] = \int_x x \cdot p\left(x|y_t\right)dx. \quad \textbf{(1)}$$

En PD-MEMLIN se proponen sendas aproximaciones tanto para x, como para la función de densidad de probabilidad, pdf, "probability density function", de x dado y_t, $p\left(x|y_t\right)$. Para ello se tienen en cuenta las siguientes consideraciones.

PD-MEMLIN asume que el espacio constituido por los vectores de características ruidosos se puede dividir en e entornos básicos, de modo que para cada uno de ellos los vectores de características se modelan como una GMM para cada fonema, ph

$$p_{e,ph}\left(y_t\right) = \sum_{s_y^{e,ph}} p\left(y_t \middle| s_y^{e,ph}\right) p\left(s_y^{e,ph}\right), \quad \textbf{(2)}$$

$$p\left(y_t \middle| s_y^{e,ph}\right) = N\left(y_t; \mu_{s_y^{e,ph}}, \Sigma_{s_y^{e,ph}}\right), \quad \textbf{(3)}$$

donde $s_y^{e,ph}$ denota la correspondiente Gaussiana del modelo de la señal ruidosa para el entorno básico e y el fonema ph, mientras que $\mu_{s_y^{e,ph}}$, $\Sigma_{s_y^{e,ph}}$ y $p\left(s_y^{e,ph}\right)$ son el vector de medias, la matriz de covarianza diagonal y la probabilidad a priori asociados a la Gaussiana $s_y^{e,ph}$.

PD-MEMLIN considera también que los distintos fonemas del espacio constituido por los vectores de características limpios se pueden modelar mediante una GMM

$$p_{ph}\left(x\right) = \sum_{s_x^{ph}} p\left(x \middle| s_x^{ph}\right) p\left(s_x^{ph}\right), \quad \textbf{(4)}$$

$$p\left(x \middle| s_x^{ph}\right) = N\left(x; \mu_{s_x^{ph}}, \Sigma_{s_x^{ph}}\right), \quad \textbf{(5)}$$

donde s_x^{ph} indica la correspondiente Gaussiana del modelo de la señal limpia para el fonema ph y $\mu_{s_x^{ph}}$, $\Sigma_{s_x^{ph}}$ y $p\left(s_x^{ph}\right)$ son el vector de medias, la matriz de covarianza diagonal y la probabilidad a priori asociados a la Gaussiana s_x^{ph}.

Finalmente, PD-MEMLIN aproxima x mediante una función lineal que depende de y_t, $s_y^{e,ph}$ y s_x^{ph}

$$x \approx \Psi\left(y_t, s_y^{e,ph}, s_x^{ph}\right) = y_t - r_{s_x^{ph}, s_y^{e,ph}}, \quad \textbf{(6)}$$

donde $r_{s_x^{ph}, s_y^{e,ph}}$ es el término independiente de la transformación asociada a las Gaussianas del espacio de la señal limpia y ruidosa del correspondiente fonema, ph. Con estas aproximaciones, la ecuación (1) se transforma en

$$\hat{x}_t = y_t - \sum_e \sum_{ph} \sum_{s_x^{ph}} \sum_{s_y^{e,ph}} p\left(e|y_t\right) p\left(ph|e, y_t\right) A_{e,ph,s_x^{ph},s_y^{e,ph}}, \quad \textbf{(7)}$$

$$A_{e,ph,s_x^{ph},s_y^{e,ph}} = p\left(s_y^{e,ph}|e, y_t, ph\right) p\left(s_x^{ph}|e, y_t, ph, s_y^{e,ph}\right) r_{s_x^{ph}, s_y^{e,ph}},$$

donde $p\left(e|y_t\right)$ es la probabilidad a posteriori del entorno básico dado el vector de características ruidoso; $p\left(ph|e, y_t\right)$ es la probabilidad a posteriori del fonema ph dado en entorno básico y el vector de características ruidoso; $p\left(s_y^{e,ph}|e, y_t, ph\right)$ es la probabilidad a posteriori de la Gaussiana $s_y^{e,ph}$ dado el entorno básico, el vector de características ruidoso y el fonema ph. Finalmente $p\left(s_x^{ph}|e, y_t, ph, s_y^{e,ph}\right)$ es la probabilidad a posteriori de la Gaussiana del modelo limpio, s_x^{ph}, dado e, y_t, ph y $s_y^{e,ph}$, o también llamada probabilidad cruzada entre Gaussianas.

2.2 Cálculo de los parámetros del estimador MMSE

A la hora de obtener la estimación del vector de características limpio, $\hat{x}_t$, se precisa calcular, tal y como queda patente en (7), distintas variables, algunas de las cuales, al depender del vector de características ruidoso, se deberán estimar durante el proceso de verificación o identificación del locutor; estamos hablando de $p\left(e|y_t\right)$, $p\left(ph|e, y_t\right)$ y $p\left(s_y^{e,ph}|e, y_t, ph\right)$. Por el contrario, el resto de variables tendrán que obtenerse previamente en una fase de entrenamiento realizada con señal estéreo: $p\left(s_x^{ph}|e, y_t, ph, s_y^{e,ph}\right)$ y $r_{s_x^{ph}, s_y^{e,ph}}$.

La probabilidad del entorno, $p(e|y_t)$, se calcula iterativamente. Para cada instante de tiempo t, $t \in [1,\ldots,T]$, se dispone de un vector de características ruidoso, y_t, de modo que el cálculo del peso del entorno básico, empleando (2) y (3), será

$$p(e|y_t) = \beta \cdot p(e|y_{t-1}) + (1-\beta)\frac{\sum_{ph} p_{e,ph}(y_t)}{\sum_{e}\sum_{ph} p_{e,ph}(y_t)}, \quad \textbf{(8)}$$

donde β es la constante de inercia del sistema y para este trabajo tendrá el valor de 0.98 ya que se puede considerar que los entornos básicos no se suceden con mucha frecuencia de una trama a otra. De cara a calcular la probabilidad del entorno para el primer vector de características, se considerará que $p(e|y_0)$ es uniforme para todos los entornos. Por otra parte, la probabilidad del fonema dado el vector de características ruidoso y el entorno, $p(ph|e,y_t)$, se calcula haciendo uso de (2) y (3) como

$$p(ph|e,y_t) = \frac{p_{e,ph}(y_t)}{\sum_{ph} p_{e,ph}(y_t)}. \quad \textbf{(9)}$$

$p(s_y^{e,ph}|e,y_t,ph)$ se puede estimar también mediante (2) y (3)

$$p(s_y^{e,ph}|e,y_t,ph) = \frac{p(y_t|s_y^{e,ph})p(s_y^{e,ph})}{\sum_{s_y^{e,ph}} p(y_t|s_y^{e,ph})p(s_y^{e,ph})}. \quad \textbf{(10)}$$

A la hora de calcular las expresiones para $p(s_x^{ph}|e,y_t,ph,s_y^{e,ph})$ y $r_{s_x^{ph},s_y^{e,ph}}$ se precisa, tal y como ya se ha indicado, de señal estéreo, de modo que se dispondrá de vectores de características limpios y ruidosos asociados a cada entorno básico y fonema: $X_{e,ph} = \{x_1^{e,ph},\ldots,x_{t_{e,ph}}^{e,ph},\ldots,x_{T_{e,ph}}^{e,ph}\}$, para la señal limpia e $Y_{e,ph} = \{y_1^{e,ph},\ldots,y_{t_{e,ph}}^{e,ph},\ldots,y_{T_{e,ph}}^{e,ph}\}$, para la señal ruidosa, con $t_{e,ph} \in [1,T_{e,ph}]$.

La probabilidad cruzada entre Gaussianas, $p\left(s_x^{ph}\middle|e, y_t, ph, s_y^{e,ph}\right)$, puede aproximarse como independiente del vector de características ruidoso, de modo que se estima como $p\left(s_x^{ph}\middle|s_y^{e,ph}\right)$, pudiéndose calcular mediante la señal estéreo usando (2), (3), (4) y (5) del siguiente modo

$$p\left(s_x^{ph}\middle|s_y^{e,ph}\right)=\frac{\sum_{t_{e,ph}} p\left(x_{t_{e,ph}}^{e,ph}\middle|s_x^{ph}\right)p\left(y_{t_{e,ph}}^{e,ph}\middle|s_y^{e,ph}\right)p\left(s_y^{e,ph}\right)p\left(s_x^{ph}\right)}{\sum_{t_{e,ph}}\sum_{s_x^{ph}} p\left(x_{t_{e,ph}}^{e,ph}\middle|s_x^{ph}\right)p\left(y_{t_{e,ph}}^{e,ph}\middle|s_y^{e,ph}\right)p\left(s_y^{e,ph}\right)p\left(s_x^{ph}\right)}. \quad (11)$$

Por su parte, el término independiente de la transformación asociada a cada par de Gaussianas del correspondiente fonema, $r_{s_x^{ph},s_y^{e,ph}}$, puede obtenerse mediante la señal estéreo minimizando el error cuadrático medio entre la señal limpia y (6) asociado para cada par de Gaussianas de cada fonema. De este modo, la expresión final obtenida será

$$r_{s_x^{ph},s_y^{e,ph}}=\frac{\sum_{t_{e,ph}} p\left(s_x^{ph}\middle|e, x_{t_{e,ph}}^{e,ph}, ph\right)p\left(s_y^{e,ph}\middle|e, y_{t_{e,ph}}^{e,ph}, ph\right)\left(y_{t_{e,ph}}^{e,ph}-x_{t_{e,ph}}^{e,ph}\right)}{\sum_{t_{e,ph}} p\left(s_x^{ph}\middle|e, x_{t_{e,ph}}^{e,ph}, ph\right)p\left(s_y^{e,ph}\middle|e, y_{t_{e,ph}}^{e,ph}, ph\right)}, \quad (12)$$

donde $p\left(s_x^{ph}\middle|e, x_{t_{e,ph}}^{e,ph}, ph\right)$ es la probabilidad de la Gaussiana del modelo de la señal limpia dado el vector de características limpio, el entorno básico y el fonema correspondientes. Dicha probabilidad se puede calcular mediante (4) y (5)

$$p\left(s_x^{ph}\middle|e, x_{t_{e,ph}}^{e,ph}, ph\right)=\frac{p\left(x_{t_{e,ph}}^{e,ph}\middle|s_x^{ph}\right)p\left(s_x^{ph}\right)}{\sum_{s_x^{ph}} p\left(x_{t_{e,ph}}^{e,ph}\middle|s_x^{ph}\right)p\left(s_x^{ph}\right)}. \quad (13)$$

El hecho de usar transformaciones independientes para cada par de Gaussianas de los distintos fonemas proporciona una normalización mucho más específica que las obtenidas por otras técnicas de normalización empíricas, como puedan ser RATZ [3], cuyas transformaciones lineales dependen de cada Gaussiana con las que se ha modelado todo el espacio limpio, SPLICE [5], en cuyo caso cada Gaussiana que modela el espacio ruidoso tiene asociada su propia transformación lineal, o incluso

MEMLIN [6], que a pesar de modelar con GMM tanto el espacio limpio como el ruidoso, no incluye la dependencia con respecto a los fonemas.

3 Los sistemas de verificación e identificación de locutor

Para la tarea de verificación de locutor se va a emplear un sistema "Universal Background Model GMM", UBM-GMM. Los vectores de características empleados en este trabajo como entrada al sistema constarán de 37 componentes y estarán constituidos por los 12 parámetros estáticos MFCC obtenidos de la señal de voz tras previa sustracción de la media del cepstrum, más la primera y segunda derivada (coeficientes delta y delta delta), además de la derivada de la energía normalizada. Se emplea además un "Voice Activity Detector", VAD, muy simple basado en energía para verificar si las distintas tramas son voz o por el contrario se trata de tramas de silencio. Cabe destacar que la duración media de las frases empleadas en verificación e identificación es de 3 segundos.

El modelo universal se entrena mediante el algoritmo "Expectation-Maximization", EM, [8] con cuatro iteraciones. Por su parte, las distintas GMMs correspondientes a los locutores se entrenan a partir del modelo universal mediante la técnica de adaptación de modelos "Maximum A Posteriori", MAP, [9].

Así pues, dada una secuencia de vectores de características de un locutor i, Y_i, el modelo universal, λ_{UBM}, y el modelo asociado al locutor i, λ_i, la decisión para determinar si es correcta la asociación entre la secuencia de tramas y el locutor vendrá dada por el siguiente cociente

$$Si \quad \frac{p(Y_i|\lambda_i)}{p(Y_i|\lambda_{UBM})} \begin{cases} < \theta \Rightarrow \lambda_i \text{ es erróneo,} \\ \geq \theta \Rightarrow \lambda_i \text{ es correcto,} \end{cases} \tag{14}$$

donde $p(Y_i|\lambda_i)$ es el valor que proporciona evaluar Y_i con el modelo λ_i; igualmente $p(Y_i|\lambda_{UBM})$ es el valor que proporciona evaluar Y_i con el modelo λ_{UBM} y finalmente θ es el umbral de decisión que se obtiene empíricamente cuando las tasas de falsos rechazos y de falsas alarmas son iguales.

El sistema de identificación de locutor se basa en los modelos (GMM) entrenados para cada locutor. Así, dada una secuencia de vectores de características Y, el sistema determinará que se trata del locutor $\hat{i}$ tal que cumpla

$$\hat{i} = \arg\max_i \left[p(Y|\lambda_i) \right]. \tag{15}$$

4 Experimentación

Para estudiar el comportamiento de los sistemas de verificación e identificación en situaciones acústicas adversas, se realizaron diversos experimentos utilizando la base de datos "Spanish SpeechDat Car" [7], que posee distintos canales (señal estéreo) y fue grabada en diferentes vehículos y condiciones de conducción. Aunque esta base de datos no se creó específicamente para las tareas de verificación e identificación de locutores, decidimos usarla por cuanto deseábamos comprobar las mejoras que se podían lograr con el método de normalización PD-MEMLIN en entornos acústicos muy variantes y complejos. Se definieron 7 entornos básicos: coche parado y motor en funcionamiento (E1), coche circulando por ciudad y condiciones no ruidosas: climatizador apagado y ventanillas subidas (E2), coche circulando por ciudad y condiciones ruidosas: ventanillas abiertas y/o climatizador encendido (E3), coche circulando a baja velocidad por pavimento en mal estado y condiciones no ruidosas (E4), coche circulando a baja velocidad por pavimento en mal estado y condiciones ruidosas (E5), coche circulando a alta velocidad por buen piso y condiciones no ruidosas (E6) y finalmente coche circulando a alta velocidad por buen piso y condiciones ruidosas (E7).

Todas las frases se muestrearon a 16 KHz. La señal limpia, a la que denominaremos en adelante como CLK ("CLose talK"), se grabó con micrófono próximo (Shure SM-10A), mientras que la ruidosa, a la que llamaremos HF ("Hands Free"), se obtuvo mediante un micrófono situado en el techo de los vehículos, encima del conductor (Peiker ME15/V520-1). El rango de SNR para la señal ruidosa va desde 4 dB hasta 14 dB, dependiendo de los entornos, mientras que la limpia cubre el intervalo desde 20 dB hasta 30 dB. Por otra parte, los parámetros MFCC estáticos se calculan cada 10 ms haciendo uso de una ventana de Hamming de 25 ms.

La técnica PD-MEMLIN se aplica sobre los 12 coeficientes MFCC estáticos más el parámetro asociado a la delta energía normalizada sin aplicar la sustracción de la media del cepstrum (ésta se eliminará tras la normalización). Para completar la descripción del sistema de normalización se mencionará que las GMMs con las que se modela cada uno de los fonemas españoles más el silencio en los espacios limpio y ruidoso se componen de 32 Gaussianas.

El modelo universal necesario en el sistema de verificación de locutor consta de 512 Gaussianas y se obtiene con el corpus de entrenamiento de la base de datos "Spanish SpeechDat Car", que está compuesto por 218 locutores y 16108 frases. Por otra parte, el corpus de evaluación de la misma base de datos consta de 91 locutores con aproximadamente 112 frases para cada una de ellos. De ellas, 50 se emplearon para entrenar las GMMs de 512 componentes asociados a cada locutor y el resto se utilizaron para estudiar el comportamiento del sistema de verificación e identificación de locutor. En las tablas 1 y 2 se pueden observar los distintos resultados, donde E1, …, E7 representan los distintos entornos básicos, EER es el "Equal error Rate", en %, CLK-CLK indica que los resultados se han obtenida empleando la señal limpia para evaluar y entrenar los distintos GMMs, HF-HF indica que es la señal ruidosa la que se utiliza tanto para evaluar los sistemas como para entrenar los GMMs, y CLK-HF norm hace referencia a que los modelos se obtuvieron con la señal limpia, mientras que las señales con las que se evaluaron son las ruidosas normalizadas previamente con PD-MEMLIN.

El número de frases para cada entorno es el siguiente: 254 para E1, 290 para E2, 235 para E3, 238 para E4, 254 para E5, 247 para E6 y 47 para E7. Además, para verificación de locutor, cada frase se compara con los 91 posibles locutores.

Tabla 1. Resultados de verificación de locutor con PD-MEMLIN para cada entorno básico.

EER	CLK-CLK	CLK-HF	HF-HF	CLK-HF norm	Mejora
E1	1.55	10.50	0.79	5.13	60.00
E2	1.21	26.73	4.70	9.58	67.20
E3	0.87	24.61	3.91	11.80	53.96
E4	0.89	27.42	2.08	6.15	70.17
E5	0.91	26.93	2.02	7.17	75.94
E6	1.08	35.00	2.71	9.71	74.56
E7	0.29	41.45	0.46	9.05	78.72
Total	1.06	26.50	3.29	8.64	70.20

Tabla 2. Resultados de identificación de locutor con PD-MEMLIN para cada entorno básico.

Tasa de acierto	E1	E2	E3	E4	E5	E6	E7	Total
CLK-CLK	99.60	99.65	99.57	99.15	100	100	100	99.69
CLK-HF	65.35	13.44	27.65	12.60	10.72	11.67	0	22.02
HF-HF	98.03	95.52	91.06	97.89	96.52	84.55	100	94.89
CLK-HF norm	86.22	61.37	49.36	68.90	44.34	56.03	48.93	59.84
Mejora	60.93	55.60	30.19	65.05	37.66	50.22	48.93	48.69

En la tabla 1 se puede observar como el ruido produce una importante degradación en el comportamiento del sistema: de aproximadamente 1% de EER, éste cae hasta 26.50%. Si la señal ruidosa se normaliza con PD-MEMLIN, se consigue una importante mejora, alcanzándose el 8.64% de tasa de falsos rechazos y falsas alarmas, lo que supone una mejora de más del 70%. Los resultados medios calculados con todos los entornos y diferentes umbrales de decisión pueden observarse en la Fig. 2.

En identificación de locutor, cuyas tasas de acierto en % pueden observarse en la tabla 2, queda igualmente patente el negativo efecto del ruido, que produce que del 99.69% de tasa de acierto en media con señal limpia se pase al 22.02% con señal ruidosa. En este caso, si se emplea PD-MEMLIN, los resultados se mejoran hasta llegar al 59.84% de tasa de acierto (48.69% de mejora).

Cabe destacar que, aunque las mejoras tras la normalización son patentes, éstas no llegan en ningún caso a los que se obtendrían si se utilizaran modelos entrenados con la señal ruidosa. Sin embargo, en muchos casos no es posible reentrenar modelos para cada locutor en todas las condiciones acústicas posibles; en ese caso las técnicas de normalización de vectores de características son una buena aproximación para acercarnos a los resultados obtenidos con señal limpia (CLK-CLK). Por otra parte, el algoritmo propuesto en este trabajo, PD-MEMLIN, proyecta del espacio ruidoso a uno limpio genérico, perdiendo parte de la especificidad propia de los locutores, por lo que se plantea como futura línea de trabajo estudiar el empleo de transformaciones

independientes para cada fonema y grupo homogéneo de locutores, “speaker clustering”.

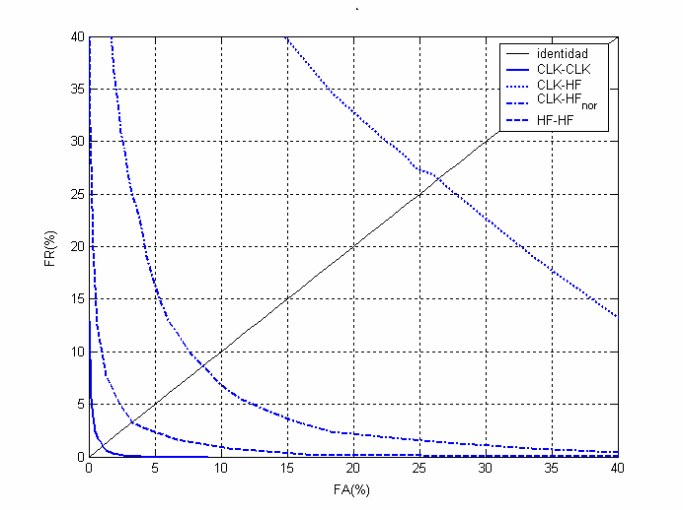

Fig. 2. Resultados de verificación para todos los entornos y diferentes umbrales de decisión.

5 Conclusiones

En este trabajo se presenta una técnica de normalización de vectores de características, PD-MEMLIN, para mejorar el comportamiento de los sistemas de verificación e identificación de locutor. El algoritmo modela cada fonema del espacio limpio y ruidoso con una GMM, además aprende una transformación lineal para cada par de Gaussianas de cada fonema. A la hora de estudiar el comportamiento de esta técnica en verificación e identificación de locutor se realizaron distintos experimentos con la base de datos “Spanish SpeechDat Car”. Se empleó un sistema UBM-GMM para verificación de locutor y un sistema GMM para identificación, obteniéndose unos resultados que muestran que el ruido produce una seria degradación en ambos sistemas, mientras que utilizando PD-MEMLIN como una fase de preprocesamiento se consigue una mejora del 70.20% en verificación y del 48.69% en identificación de locutor. Como futura línea de trabajo se propone la utilización de transformaciones lineales dependientes de grupos de locutores afines por cuanto la proyección sobre un espacio limpio genérico hace perder parte de la especificidad de los distintos locutores.

Referencias

1. Reynold D. A.., Quatieri T. F., Dunn R. B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, (2000) vol 10, pp 19-41.
2. Stern R. M., Raj B., Moreno P. J.: Compensation for environmental degradation in automatic speech recognition: in Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, (1997), pp. 33-42.
3. Moreno P. J.: Speech Recognition in Noisy Environments, Ph.D. Theses, ECE Department, Carnegie-Mellon University (1996).
4. Acero A.: Acustical and Environmental Robustness in Automatic Speech Recognition, Ph.D. Theses, ECE Department, Carnegie-Mellon University (1990).
5. Droppo J. Deng L. Acero A.: Evaluation of the SPLICE Algorithm on the Aurora2 Database, in Proc. Eurospeech, (2001).
6. Buera L. Lleida E. Miguel A. Ortega A.: Multi-Environment Models Based Linear Normalization for Speech Recognition in Car Conditions, in Proc. ICASSP, (2004).
7. Moreno A., Nogueira A.: SpeechDat-Car: Spanish, Technical report SpeechDat.
8. Bilmes J.: A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical report, University of Berkeley, ICSI-TR-97-021 (1991).
9. Gauvain J.-L., Lee C.-H.: Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, IEEE Trans. On Speech and Audio Processing, (1994), vol. 2, pp. 291-298.