Avances en la Normalización Cepstral con Señal Estéreo para el Reconocimiento Robusto de Voz en el Entorno del Vehículo

Luis Buera, Eduardo Lleida, Antonio Miguel, Alfonso Ortega y Óscar Saz

Universidad de Zaragoza.
{lbuera, lleida, amiguel, ortega, oskarsaz}@unizar.es

Resumen

Las técnicas clásicas de normalización cepstral tienen dos importantes problemas. El primero es la suposición de independencia entre los coeficientes de los vectores de características, que hace que ciertos efectos producidos por el ruido, como las rotaciones espaciales, no puedan ser compensados. El segundo problema es el desajuste que hay entre las transformaciones propuestas para la normalización y las ideales que se deberían aplicar. En este artículo se proponen dos aproximaciones para solucionar estos problemas: una transformación de rotación con varios entornos y el uso de modelos acústicos del espacio transformado. Se realizaron varios experimentos con la base de datos SpeechDat Car para estudiar el comportamiento de estas dos técnicas, probándolos con los algoritmos de normalización "Multi-Environment Models based LInear Normalization"(MEMLIN) y "Multi-Environment Models based HIstogram Normalization"(MEMHIN), obteniéndose una mejora media del "Word Error Rate"(WER) del 83.62 %.

1. Introducción

Cuando las condiciones acústicas entre el test y el entrenamiento son diferentes, las prestaciones de los sistemas de reconocimiento del habla caen estrepitosamente. Para compensarlo, durante los últimos años se han desarrollado diferentes técnicas. Éstas pueden clasificarse en dos grandes grupos [1]: adaptación de modelos acústicos y compensación de los vectores de características, o normalización. El primero de los grupos, que sólo modifica los modelos acústicos, proporciona, en principio, transformaciones más específicas, mientras que las técnicas de normalización, que modifican únicamente los vectores de características, necesitan menos datos y tiempo computacional. Por otra parte, existen algoritmos híbridos, que también se han mostrado efectivos [2]. Debido a sus distintas características, el uso de una u otra familia de técnicas dependerá finalmente de la aplicación que se vaya a desarrollar.

A su vez, las técnicas de normalización pueden dividirse en tres grandes grupos [3]: compensación basada en modelos, compensación empírica y filtrado cepstral paso alto. El primero de ellos consiste en suponer que la contaminación de la señal viene dada por un determinado entorno, cuyos parámetros se estiman mediante la señal sucia que se posee. El segundo grupo no hace suposición alguna y se basa en comparar la señal limpia y sucia para determinar las transformaciones que se deben llevar a cabo en la normalización. En este caso se precisa bases de datos estéreo. Por último, el filtrado cepstral, aun produciendo peores resultados que las técnicas anteriores, es muy utilizado debido a que su coste computacional es prácticamente nulo.

En este artículo, se presentan dos mejoras para hacer más robustas las técnicas clásicas de normalización. La primera consiste en compensar mediante varios entornos la rotación espacial que el ruido puede introducir. Este algoritmo compensa uno de los efectos que la presunción de independencia entre los coeficientes de los vectores de características no puede corregir. La segunda técnica propuesta consiste en usar modelos acústicos del espacio transformado a la hora de reconocer. De este modo se reduciría el efecto del error de desajuste entre las transformaciones ideales y las que realmente se emplean. Estas dos mejoras propuestas has sido probadas con MEMLIN y MEMHIN, algoritmos estos que ya se ha constatado que compensan adecuadamente en [4] y [5] .

Este artículo se organiza del siguiente modo: en la sección 2 se realiza un breve repaso de las técnicas de normalización basadas en el mínimo error cuadrático medio, MMSE. La transformación de rotación multi-entorno se introduce en el apartado 3. La estrategia del uso de modelos acústicos del espacio transformado se incluye en la siguiente sección. Los resultados con la base de datos SpeechDat Car [6] se representan y discuten en la sección 5. Finalmente se incluyen las conclusiones en la sección 6.

2. Estimador MMSE

Las técnicas de normalización basadas en MMSE han tenido en los últimos tiempos un gran auge, surgiendo múltiples algoritmos. Tal y como se va a poder comprobar a continuación, todos éstos parten del mismo desarrollo teórico.

El vector de características estimado, $\hat{x}$, se puede obtener mediante la estimación de mínimo error cuadrático medio del siguiente modo:

$$\hat{x} = E[x|y] = \int_x xp(x|y)dx \tag{1}$$

donde $p(x|y)$ es la función de densidad de probabilidad, PDF, de x, dado y.

A la hora de calcular la expresión anterior, se requieren una serie de aproximaciones. La elección de unas u otras determinará el algoritmo en cuestión.

2.1. Aproximaciones

En general, se pueden asumir tres aproximaciones: primera, el espacio sucio se divide en varios entornos básicos y los vectores de características sucios, y, se pueden modelar estadísticamente mediante una mezcla de gaussianas para cada uno de dichos entornos básicos:

$$p_e(y) = \sum_{s_y^e} p(y|s_y^e)p(s_y^e) \tag{2}$$

$$p(y|s_y^e) = N(y; \mu_{s_y^e}, \Sigma_{s_y^e}) \tag{3}$$

donde s_y^e indica la correspondiente gaussiana del modelo sucio para el entorno e; $\mu_{s_y^e}$, $\Sigma_{s_y^e}$ y $p(s_y^e)$ son el vector de media, la matriz diagonal de covarianzas y el peso asociado a s_y^e.

Segunda, los vectores de características limpios, x, siguen la distribución de una mezcla de gaussianas:

$$p(x) = \sum_{s_x} p(x|s_x)p(s_x) \tag{4}$$

$$p(x|s_x) = N(x; \mu_{s_x}, \Sigma_{s_x}) \tag{5}$$

donde s_x es la correspondiente gaussiana del modelo limpio; μ_{s_x}, Σ_{s_x} y $p(y|s_x)$ son el vector de medias, la matriz diagonal de covarianzas y el peso asociado a s_x.

Tercera, se considera que x puede ser aproximado por una función del vector de características sucio, y, la gaussiana del modelo limpio, s_x, y la gaussiana del modelo sucio, s_y^e:

$$x \simeq f(y, s_x, s_y^e) \tag{6}$$

En el caso de "multivariate gaussian based cepstral normalization", RATZ, [7] las aproximaciones se reducen a suponer que la señal limpia se puede modelar como una mezcla de gaussianas (4) y (5) y a establecer la siguiente función de estimación para el vector de características limpio:

$$x \simeq f_{RATZ}(y, s_x, s_y^e) = f_{RATZ}(y, s_x) = y - r_{s_x} \tag{7}$$

donde r_{s_x} es el término independiente de la transformación lineal propuesta para el algoritmo RATZ.

Para el caso de "Stereo based Piecewise LInear Compensation for Environments", SPLICE, [8], las aproximaciones serán la suposición de que la señal sucia se puede modelar mediante una mezcla de gaussianas (pero sin dividir el espacio sucio en entornos básicos) y la utilización de la siguiente función de estimación para el vector x:

$$x \simeq f_{SPLICE}(y, s_x, s_y^e) = f_{SPLICE}(y, s_y) = y - r_{s_y} \tag{8}$$

donde r_{s_y}, que depende de la gaussiana sucia, es el término independiente de la transformación lineal propuesta para el algoritmo SPLICE.

Tanto MEMLIN como MEMHIN dividen el espacio sucio en varios entornos básicos y, para cada uno de ellos, se supone que se pueden modelar mediante una mezcla de gaussianas. Del mismo modo, el espacio limpio se puede representar como una mezcla de gaussianas. En cuanto a las funciones de estimación, éstas sí que varían según el algoritmo:

$$x \simeq f_{MEMLIN}(y, s_x, s_y^e) = y - r_{s_x, s_y^e} \tag{9}$$

$$x \simeq f_{MEMHIN}(y, s_x, s_y^e) = C^{-1}_{x, s_x, s_y^e}(C_{y, s_x, s_y^e}(y)) \tag{10}$$

donde r_{s_x, s_y^e} es el término independiente de la transformación lineal de la técnica MEMLIN, que dependerá de cada par de gaussianas, s_x y s_y^e. En la función correspondiente a MEMHIN, C^{-1}_{x, s_x, s_y^e} es la inversa de la función de probabilidad acumulada de la señal limpia asociada a las gaussianas s_x y s_y^e, mientras que C_{y, s_x, s_y^e} es la función de probabilidad acumulada de los vectores de características sucios asociada a las gaussianas s_x y s_y^e.

2.2. Expresiones finales

Mediante la aproximaciones planteadas en el apartado anterior, la ecuación (1), se puede aproximar para los algoritmos RATZ (11), SPLICE (12), MEMLIN (13), y MEMHIN (14) del siguiente modo:

$$\hat{x}_t \simeq y_t - \sum_{s_x} r_{s_x} p(s_x|y_t) \tag{11}$$

$$\hat{x}_t \simeq y_t - \sum_{s_y} r_{s_y} p(s_y|y_t) \tag{12}$$

$$\hat{x}_t \simeq y_t - \sum_{s_x} \sum_{e} \sum_{s_y^e} \alpha_{e,t} r_{s_x, s_y^e} p(s_y^e|y_t, e) p(s_x|s_y^e, y_t) \tag{13}$$

$$\hat{x}_t \simeq \sum_{s_x} \sum_{e} \sum_{s_y^e} \alpha_{e,t} p(s_y^e|y_t, e) p(s_x|s_y^e, y_t) C^{-1}_{x, s_x, s_y^e}(C_{y, s_x, s_y^e}(y_t)) \tag{14}$$

donde t es el índice temporal. $p(s_x|y_t)$ es la probabilidad de la gaussiana limpia dada la trama sucia y_t. $p(s_y|y_t)$ es la probabilidad de la gaussiana sucia dada y_t. $\alpha_{e,t}$ es el peso del entorno dado y_t. $p(s_y^e|y_t, e)$ es la probabilidad de la gaussiana sucia dado el vector de características sucio y el entorno. Finalmente, $p(s_x|s_y^e, y_t)$ es la probabilidad condicionada de la gaussiana limpia dados s_y^e e y_t.

2.3. Cálculo de los parámetros

A la hora de calcular la estimación del vector limpio, $\hat{x}_t$, se deben obtener una serie de variables, a saber: $p(s_x|y_t)$, $p(s_y|y_t)$, $\alpha_{e,t}$, $p(s_y^e|y_t, e)$, $p(s_x|s_y^e, y_t)$, r_{s_x, s_y^e}, r_{s_x}, r_{s_y}, C_{x, s_x, s_y^e} y C_{y, s_x, s_y^e}. Las primeras cuatro, que dependen de vector sucio de características que se pretende normalizar, se calculan durante el reconocimiento. El resto de variables se obtendrán mediante un proceso de entrenamiento durante el que se precisa señal estéreo.

Para calcular el peso del entorno, $\alpha_{e,t}$, se utiliza una solución iterativa. Para cada instante de tiempo, t, se dispone de un vector de características sucio, y_t, de modo que la expresión de $\alpha_{e,t}$:

$$\alpha_{e,t} = \beta \cdot \alpha_{e,t-1} + (1 - \beta) \frac{p_e(y_t)}{\sum_e p_e(y_t)} \tag{15}$$

donde β es la constante de memoria. Se considerará que $\alpha_{e,0}$ es uniforme para todos los entornos.

$p(s_y^e|y_t, e)$ se calculará mediante (2), (3) y el teorema de Bayes:

$$p(s_y^e|y_t, e) = \frac{p(y_t|s_y^e)p(s_y^e)}{\sum_{s_y^e} p(y_t|s_y^e)p(s_y^e)} \tag{16}$$

$p(s_y|y_t)$ y $p(s_x|y_t)$ se obtienen de un modo semejante al anterior.

A la hora de obtener las expresiones para $p(s_x|s_y^e, y_t)$, r_{s_x, s_y^e}, r_{s_x}, r_{s_y}, C_{x, s_x, s_y^e} y C_{y, s_x, s_y^e}, se precisa de un proceso de entrenamiento en el que se hará uso de señal estéreo para cada entorno que, de un modo general se puede expresar del siguiente modo: $X_e = \{x_1^e, ..., x_{t_e}^e, ..., x_{T_e}^e\}$, para los vectores de características limpios e $Y_e = \{y_1^e, ..., y_{t_e}^e, ..., y_{T_e}^e\}$ para los sucios, donde $t_e \in [1, T_e]$. Para el caso de RATZ y SPLICE no hay entornos alguno y, por tanto, el índice e carece de sentido.

	E1	E2	E3	E4	E5	E6	E7
C0-C2	20.49	17.62	23.19	22.41	21.27	21.17	25.77
C0-MEMLIN 32-32	5.62	5.44	5.96	7.45	6.53	6.99	8.48
C0-MEMLIN 32-32 Rot	0.023	0.029	0.020	0.039	0.032	0.032	0.029

Tabla 1: Ángulos en grados entre los ejes de máxima varianza para siete condiciones de conducción (E1... E7).

La probabilidad condicionada, $p(s_x|s_y^e, y_t)$, se puede considerar independiente del tiempo y se estima mediante frecuencia relativa:

$$p(s_x|s_y^e, y_t) \simeq p(s_x|s_y^e) = \frac{C_N(s_x|s_y^e)}{N} \quad (17)$$

donde $C_N(s_x|s_y^e)$ es el número de veces en que, para cada par de vectores de características del entorno e, el par de gaussianas más probable es s_x y s_y^e. N es el número de veces en que la gaussiana más probable en el entorno e para el vector sucio es s_y^e.

Para el algoritmo MEMLIN, el cálculo de la variable r_{s_x,s_y^e} se realiza mediante la minimización del error cuadrático, E_{s_x,s_y^e}:

$$E_{s_x,s_y^e} = \sum_{t_e} p(s_x|x_{t_e})p(s_y^e|y_{t_e})(x_{t_e} - y_{t_e} + r_{s_x,s_y^e})^2 \quad (18)$$

$$r_{s_x,s_y^e} = \frac{\sum_{t_e} p(s_x|x_{t_e}^e)p(s_y^e|y_{t_e}^e)(y_{t_e}^e - x_{t_e}^e)}{\sum_{t_e} p(s_x|x_{t_e}^e)p(s_y^e|y_{t_e}^e)} \quad (19)$$

donde $p(s_x|x_{t_e}^e)$ es la probabilidad de s_x, dado el vector de características limpio, y se puede calcular mediante (4), (5) y el teorema de Bayes del mismo modo que (16):

$$p(s_x|x_{t_e}^e) = \frac{p(x_{t_e}|s_x)p(s_x)}{\sum_{s_x} p(x_{t_e}|s_x)p(s_x)} \quad (20)$$

El cálculo de r_{s_x} y r_{s_y} para RATZ y SPLICE se hace de un modo similar al del MEMLIN. De este modo, las expresiones finales serán, respectivamente:

$$r_{s_x} = \frac{\sum_t p(s_x|x_t)(y_t - x_t)}{\sum_t p(s_x|x_t)} \quad (21)$$

$$r_{s_y} = \frac{\sum_t p(s_y|y_t)(y_t - x_t)}{\sum_t p(s_y|y_t)} \quad (22)$$

Por otra parte, para el algoritmo MEMHIN y asociado a cada par de gaussianas (una del modelo limpio, otra del modelo sucio del correspondiente entorno), se realizan los histogramas con n bandas para cada coeficiente del vector de características (dichos coeficientes se consideran independientes [8]). A la hora de obtener los histogramas, las tramas se pesan por $p(s_x|x_{t_e})p(s_y^e|y_{t_e})$. Una vez obtenidos, las funciones de probabilidad acumuladas, C_{x,s_x,s_y^e} y C_{y,s_x,s_y^e}, se calculan directamente.

3. Transformación de rotación multi-entorno

En general, las técnicas clásicas de normalización suponen que los coeficientes de los vectores de características son independientes. De este modo, algunos de los efectos que produce el ruido no pueden ser compensados, como las rotaciones espaciales. Esto queda patente en la tabla 1, donde C0-C2 indica el ángulo que hay entre los ejes de máxima varianza de los vectores de características limpios (C0) y los sucios (C2). C0-MEMLIN 32-32 representa el ángulo que hay entre la señal limpia y la normalizada cuando se ha aplicado el algoritmo MEMLIN con 32 gaussianas para los modelos limpio y sucios. Los resultados muestran que la técnica de normalización aplicada no es suficiente para compensar toda la rotación que los distintos entornos han introducido, pero que si se aplica la técnica de transformación de rotación multi-entorno tras MEMLIN, los ángulos llegan a ser prácticamente nulos (C0-MEMLIN 32-32 Rot). Los resultados aplicando MEMHIN son muy similares.

El principio de la técnica de la rotación [9] es obtener una matriz de transformación (U_1) para normalizar los vectores de características. El subíndice 1 indica que únicamente se va a modificar la dirección de los ejes de máxima varianza y no el resto.

$$\hat{x}_t = U_1 y_t \quad (23)$$

Mediante el corpus de train estéreo de la base de datos, se puede obtener una matriz de transformación para cada entorno, $U_{e,1}$. A la hora de determinar los ejes de máxima varianza en los espacios limpio y sucio, se utilizará el análisis de componentes principales (PCA). Este análisis se realizará sobre las matrices de covarianzas de los vectores de características limpios y sucios para cada entorno, ($\tilde{\Sigma}_e$, Σ_e, respectivamente). Los correspondientes vectores y valores propios serán: $\tilde{v}_{e,i}$ y $\tilde{\lambda}_{e,i}$, para el espacio limpio, y $v_{e,i}$ y $\lambda_{e,i}$, para el sucio, donde $i = 1...D$, $\tilde{\lambda}_{e,1} \geq \tilde{\lambda}_{e,2} \geq ... \geq \tilde{\lambda}_{e,D}$, $\lambda_{e,1} \geq \lambda_{e,2} \geq ... \geq \lambda_{e,D}$ y D es la dimensión de los vectores de características. El ángulo de rotación entre las dos principales direcciones de los espacios limpio y sucio se obtiene del siguiente modo: $\eta_{e,1} = \arccos(\tilde{v}_{e,1} \cdot v_{e,1})$. Por otra parte, se puede considerar que $\tilde{v}_{e,1}$ y $v_{e,1}$ determinan un plano ($\pi_{e,1}$). La idea geométrica que se encuentra bajo esta transformación es dividir cada vector de características en dos partes: la proyección sobre $\pi_{e,1}$, que se rotará $\eta_{e,1}$ grados, y la parte perpendicular al plano, que no se modificará.

Como $\tilde{v}_{e,1}$ y $v_{e,1}$ no son ortogonales, se aplica Gram-Schmidt a $v_{e,1}$ para obtener una base de vectores ortonormales: $\hat{v}_{e,1}$ y el propio $\tilde{v}_{e,1}$, que definan el mismo plano de rotación, $\pi_{e,1}$:

$$\hat{v}_{e,1} = \frac{v_{e,1} - (\tilde{v}_{e,1} \cdot v_{e,1}) \cdot \tilde{v}_{e,1}}{\|v_{e,1} - (\tilde{v}_{e,1} \cdot v_{e,1}) \cdot \tilde{v}_{e,1}\|} \quad (24)$$

$J_{e,1}^T$ es la matriz de proyección sobre $\pi_{e,1}$ y $R_{e,1}$ es la transformación de la rotación para el ángulo $\eta_{e,1}$:

$$J_{e,1}^T = (\hat{v}_{e,1}, \tilde{v}_{e,1})^T \quad (25)$$

$$R_{e,1} = \begin{pmatrix} \cos(\eta_{e,1}) & -\sin(\eta_{e,1}) \\ \sin(\eta_{e,1}) & \cos(\eta_{e,1}) \end{pmatrix} \quad (26)$$

Train	Test	E1	E2	E3	E4	E5	E6	E7	MWER	MIMP
C0	C0	1.90	2.64	1.81	1.75	1.62	0.64	0.35	1.75	
C0	C2	5.91	14.49	14.55	20.17	21.07	16.19	35.71	16.21	
C2	C2	6.67	14.24	12.73	12.91	14.97	9.68	8.50	11.81	
C2†	C2	2.86	7.12	4.34	4.39	7.63	4.60	4.76	5.30	
MEMLIN 8-8		3.16	8.49	6.43	9.27	11.91	9.05	14.97	8.39	56.00
MEMLIN 16-16		3.26	8.06	5.45	7.64	10.01	7.78	12.92	7.37	61.49
MEMLIN 32-32		2.49	7.80	5.03	6.64	9.25	6.51	11.22	6.62	68.50
MEMHIN 8-8		3.07	7.98	6.57	8.02	11.63	7.78	12.59	7.80	59.62
MEMHIN 16-16		2.97	7.98	6.15	7.14	9.91	7.76	11.56	7.20	63.17
MEMHIN 32-32		2.40	7.80	5.31	6.52	8.77	6.51	8.50	6.39	69.58

Tabla 2: Resultados de referencia de tasa de error de palabras. C2†significa que el entrenamiento para cada entorno es específico

Finalmente, la matriz de transformación para el correspondiente entorno, $U_{e,1}$, se puede obtener como:

$$U_{e,1} = J_{e,1} R_{e,1} J_{e,1}^T + I + J_{e,1} J_{e,1}^T \quad (27)$$

donde I es la matriz identidad. La transformación de la rotación se podría realizar sobre todos los ejes que proporciona el análisis PCA y no sólo sobre el de máxima varianza, pero se ha demostrado que con el primer vector es suficiente [9]. En reconocimiento, todas las tramas de cada frase se normalizan mediante la matriz de transformación del entorno más probable, $\hat{e}$. Esto es: $U_1 = U_{\hat{e},1}$.

4. Modelos acústicos del espacio transformado

En general, las técnicas de normalización proyectan los vectores de características sobre el espacio limpio mediante la correspondiente transformación. Dado que estas transformaciones no son perfectas, el nuevo espacio transformado que se obtiene no es el limpio, tal y como debería ser en el caso ideal. Este error de desajuste puede solucionarse en reconocimiento mediante el aprendizaje de nuevos modelos acústicos, que se obtendrían mediante la proyección del corpus de train sucio de la base de datos desde el propio espacio sucio hacia el limpio. A estos modelos se les denomina modelos acústicos del espacio transformado.

En reconocimiento, a la hora de normalizar las señales de test, se tienen dos opciones: la primera de ellas consistiría en usar las mismas transformaciones empleadas para normalizar la señal de train, esto es, proyectar desde el espacio sucio al limpio. La segunda opción consiste en utilizar unas nuevas transformaciones que proyectarían desde el espacio sucio al transformado, que viene definido por la señal de train normalizada del modo que se ha comentado anteriormente.

En este artículo, el corpus sucio de train se normaliza mediante MEMLIN o MEMHIN; del mismo modo, las transformaciones entre el espacio sucio y el transformado también se obtendrán mediante las mismas técnicas.

5. Resultados

Se llevaron a cabo una serie de experimentos usando la base de datos SpeechDat Car en español [6]. Se definieron siete entornos básicos que se corresponden con diversas situaciones de conducción: coche parado y motor en funcionamiento (E1), conducción por ciudad con las ventanillas subidas y el climatizador apagado (condiciones silenciosas) (E2), conducción por ciudad con condiciones ruidosas: ventanillas abiertas y/o climatizador encendido (E3), conducción a baja velocidad por mala carretera y en condiciones silenciosas (E4), conducción a baja velocidad por mala carretera y en condiciones ruidosas (E5), alta velocidad, buen piso y condiciones silenciosas (E6) y, por último, conducción a alta velocidad por buena carretera y bajo condiciones ruidosas (E7).

La tarea que se empleó para los experimentos fue la de dígitos, tanto continuos como aislados. Las frases fueron grabadas muestreando a 16 KHz. Las señales limpias fueron grabadas mediante un micrófono próximo (Shune SM-10A), al que a partir de ahora se le llamará C0. Por otra parte, las señales sucias se grabaron mediante un micrófono situado en el techo del vehículo encima del conductor (Peiker ME15/V520-1), que se le denominará C2. El rango de SNR para la señal limpia va desde 20 hasta 30 dB, mientras que para la señal ruidosa va desde 5 a 20 dB. Por otra parte, los 12 coeficientes Mel Cepstrum, MFCC, y la energía se obtienen cada 10 ms usando una ventana de Hamming de 25 ms.

Las técnicas de normalización propuestas en este artículo se aplican sobre los 12 coeficientes MFCC y la derivada de la energía. Por otra parte, los modelos que se emplean podrán ser de 8, 16 ó 32 gaussianas.

En reconocimiento, el vector de características se compone de los doce coeficientes MFCC, sobre los que se ha aplicado la sustracción de la media del cepstrum, la primera y la segunda derivadas, así como la derivada de la energía, constituyendo pues un vector de 37 coeficientes. Por otra parte, los modelos acústicos son fonéticos, de modo que habrá 25 HMM continuos de tres estados y 16 gaussianas por estado. Cada uno de ellos modelará un fonema. También habrá dos modelos de silencio, uno para el silencio largo y otro para modelar el que se produce entre palabras.

Los resultados de referencia para cada entorno se presentan en la tabla 2. En ella, junto a la técnica utilizada aparece el número de las gaussianas de los modelos limpio y sucios, respectivamente. MWER representa la tasa de error media, obtenida proporcionalmente al número de frases de cada entorno mediante las correspondientes tasas de error. MIMP es la media de la mejora que se obtiene entre la primera y la segunda fila y se calcula del mismo modo que MWER. Se puede

	MWER	MIMP
MEMLIN 8-8 Rot	7.89	59.70
MEMLIN 16-16 Rot	7.01	64.42
MEMLIN 32-32 Rot	6.20	70.61
MEMHIN 8-8 Rot	7.27	62.07
MEMHIN 16-16 Rot	6.50	67.52
MEMHIN 32-32 Rot	5.99	71.68
MEMLIN 8-8 Rtr	4.64	79.43
MEMLIN 16-16 Rtr	4.36	82.42
MEMLIN 32-32 Rtr	4.20	82.24
MEMHIN 8-8 Rtr	4.64	80.23
MEMHIN 16-16 Rtr	4.32	81.38
MEMHIN 32-32 Rtr	4.25	81.32
MEMLIN 8-8 Rot Rtr	4.83	77.63
MEMLIN 16-16 Rot Rtr	4.52	81.20
MEMLIN 32-32 Rot Rtr	4.29	82.58
MEMHIN 8-8 Rot Rtr	4.71	79.95
MEMHIN 16-16 Rot Rtr	4.35	80.83
MEMHIN 32-32 Rot Rtr	4.20	82.47

Tabla 3: Tasas medias de WER y mejoras para las mejoras propuestas aplicadas a MEMLIN y MEMHIN.

	MWER	MIMP
MEMLIN 16-16. MEMLIN 8-8	4.76	79.96
MEMLIN 16-16. MEMLIN 16-16	4.46	82.12
MEMLIN 16-16. MEMLIN 32-32	4.55	80.69
MEMHIN 16-16. MEMHIN 8-8	4.62	80.62
MEMHIN 16-16. MEMHIN 16-16	4.27	82.47
MEMHIN 16-16. MEMHIN 32-32	4.23	83.62

Tabla 4: Tasas medias de WER y mejoras cuando se utilizan modelos acústicos del especio transformado y proyectando en reconocimiento desde el espacio sucio al transformado.

apreciar que los resultados de MEMHIN son mejores que los que se obtienen mediante la aplicación de la técnica MEMLIN, especialmente cuando los entornos son más ruidosos: esto es así porque en esos entornos el ruido aditivo es más importante y MEMHIN lo compensa de un modo más eficiente [5].

Los resultados comparativos entre las diferentes mejoras propuestas se encuentran en las tablas 3 y 4.

En la tabla 3 se presentan los resultados cuando se aplican las mejoras de la transformación de rotación multi-entorno (Rot en la tabla) y del uso de modelos acústicos del espacio transformado (Rtr). En todos los casos, la proyección en reconocimiento se hará entre el espacio sucio y el limpio. La posición en el que aparecen los distintos nombres de las técnicas en la tabla indica el orden en que se aplicaron.

Se puede observar que al añadir la transformación de rotación multi-entorno después de MEMLIN o MEMHIN, se produce una ligera mejora con respecto a los resultados que ya de por sí eran capaces de obtener ambas técnicas de normalización (hasta un 4.35 % de mejora en WER). Esto muestra que la mejora en los ángulos de los ejes de máxima varianza (tabla 1) tienen asociado también una mejora en la tasa de reconocimiento. Usar los modelos acústicos del espacio transformado produce importantes mejoras (hasta un 23.49 % en WER entre MEMLIN 8-8 y MEMLIN 8-8 Rtr). La combinación de las dos técnicas proporciona el mejor resultado (82.58 % en WER). Otra ventaja del uso de los modelos acústicos del espacio transformado es que el comportamiento en reconocimiento es casi independiente del número de gaussianas de los modelos que se emplean en MEMLIN o MEMHIN.

En la tabla 4, la estrategia de usar los modelos acústicos del espacio transformado es aplicada a MEMLIN y MEMHIN, utilizando en reconocimiento la proyección entre el espacio sucio y el transformado. Los nuevos modelos acústicos se calcularon al aplicar el primer algoritmo que aparece en la tabla sobre el corpus de train sucio de la base de datos (proyectando del espacio sucio al limpio). La segunda técnica que aparece en la tabla se usa para normalizar en reconocimiento, proyectando desde el espacio sucio al transformado. Se puede observar que la mejora (83.62 % en WER) es mejor que las obtenidas con otras técnicas. Se obtuvieron los resultados para otras combinaciones de gaussianas y los resultados fueron muy similares. Por otra parte, y aunque no se hayan presentado los resultados para cada uno de los entornos, se puede constatar como el uso de los modelos acústicos del espacio transformado proporciona mejores resultados para todos los entornos que los que se llegan a obtener mediante el entrenamiento específico ($C2\dagger$).

6. Conclusiones

En este artículo se han presentado dos mejoras para las técnicas clásicas de normalización, obteniéndose importantes mejoras al combinarlas con los algoritmos MEMLIN y MEMHIN. La transformación de rotación multi-entorno, que compensa algún problema de la aproximación que considera que los coeficientes de los vectores de características son independientes, produce una mejora del 4.35 % en WER respecto a MEMLIN y MEMHIN. Usar los modelos acústicos del espacio transformado y aplicar en reconocimiento una proyección del espacio sucio al limpio, logra una mejora de más del 11.74 % en WER. Combinando ambas técnicas se puede llegar a lograr hasta una mejora del 82.58 % en WER. El uso de los modelos acústicos del espacio transformado, proyectando en reconocimiento desde el espacio sucio al transformado obtiene los mejores resultados (hasta 83.62 % en WER). Por otra parte, los nuevos modelos acústicos hacen que las tasas de reconocimiento sean menos sensibles al número de gaussianas que se utilicen.

7. Referencias

[1] A. Acero, y X. Huang, "Augmented Cepstral Normalization for Robust Speech Recognition", Proc. of IEEE Automatic Speech Recognition Workshop, pp 146–147, Dec. 1995.

[2] A. Sankar, y C. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", IEEE Trans. on Speech and Audio Processing, pp 190–202, May. 1996.

[3] R. M. Stern, B. Raj, y P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition", in Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, pp 33–42, Apr. 1997.

[4] L. Buera, E. Lleida, A. Miguel, y A. Ortega, ”Multi-environment models based linear normalization for speech recognition in car conditions”, in Proc. ICASSP, May. 2004.

[5] L. Buera, E. Lleida, A. Miguel, y A. Ortega, ”Multi-environment models based linear normalization for robust speech recognition”, in Proc. SPECOM, Sep. 2004.

[6] A. Moreno, A. Noguiera, A. Sesma, ”SpeechDat-Car: Spanish”, Technical Report SpeechDat.

[7] P. Moreno, ”Speech recognition in noisy environments”, Ph. D. Thesis, ECE Departament, Carnegie-Mellon University, Apr. 1996.

[8] J. Droppo, L. Deng, y A. Acero, ”Evaluation of the SPLICE algorithm on the Aurora2 database”, in Proc. Eurospeech, vol. 1, Sep. 2001.

[8] Ángel de la Torre, Antonio M. Peinado, Jose C. Segura, Jose L. Pérez, Carmen Benítez, y Antonio J. Rubio, ”Histogram equalization of the speech representation for robust speech recognition”, IEEE Trans. on Speech and Audio Processing, to be published. 2004.

[9] S.Molau, ”Normalization in the Acoustic Feature Space for Improved Speech Recognition”, Ph. D. Thesis, Computer Science Department, RWTH Aachen. Feb. 2003.