RECONOCIMIENTO DEL LOCUTOR MEDIANTE FILTRADO FRECUENCIAL DE ENERGÍAS ESPECTRALES ESTIMADAS POR MÉTODOS HÍBRIDOS*

Javier Hernando, Climent Nadeu

Centro TALP, Depto. TSC, UPC, Barcelona, javier@gps.tsc.upc.es

RESUMEN

Se han explorado dos formas de obtener parámetros más robustos para reconocimiento del locutor: la hibridación de técnicas de análisis espectral y el filtrado frecuencial de las energías de las bandas. Se ha comprobado que el filtrado frecuencial constituye una representación eficiente en reconocimiento del habla y puede ecualizar aproximadamente la varianza cepstral, realzando las oscilaciones espectrales más efectivas para la discriminación entre locutores. Se han obtenido buenos resultados de identificación sobre la base de datos TIMIT, especialmente cuando se ha añadido ruido blanco. Por otro lado, se ha explorado la hibridación de la predicción lineal y el banco de filtros en la etapa de análisis espectral. La combinación de estas técnicas ha proporcionado buenos resultados de verificación sobre la base de datos telefónica POLYCOST.

1. INTRODUCCIÓN

En los sistemas actuales de reconocimiento automático de locutores, la envolvente espectral de cada tramo de voz suele representarse mediante un conjunto de coeficientes de la serie de Fourier de su logaritmo, es decir, los coeficientes cepstrales C(m), $1 \le m \le M$. Estos parámetros tienen un alto grado de especificidad respecto al locutor [1]. Suelen calcularse a partir de una estimación espectral de predicción lineal (LP), LP-cepstrum, o de las energías de un banco de filtros (FB) espaciados en la escala mel (mel-cepstrum). Sin embargo, hay pocos estudios comparativos sobre la robustez relativa al ruido y las distorsiones de ambos métodos.

Recientemente, los autores han considerado un esquema unificado de parametrización para reconocimiento del habla que combina ambos tipos de análisis [2]. Se ha comprobado que una hibridación apropiada de ambas aproximaciones puede superar los resultados de reconocimiento en reconocimiento de dígitos tanto en ausencia como en presencia de ruido.

Por otro lado, podemos preguntarnos si los coeficientes cepstrales son el mejor modo de representar la envolvente espectral. La secuencia de coeficientes cepstrales C(m) es una representación compacta y cuasi-incorrelada de los espectros de voz y siempre se enventana antes del cálculo de distancia o probabilidad en la etapa de ajuste de patrones del proceso de reconocimiento. Esta ventana elimina los coeficientes cepstrales a partir de la cuefrencia M y, en algunos tipos de sistemas, pondera adecuadamente los coeficientes restantes [1] [3] [4] [5] para aumentar la capacidad de discriminación del sistema.

* Este trabajo ha sido subvencinado por las proyectos TIC 98-0685 y TIC 98-0423-C06-01

Sin embargo, los coeficientes cepstrales tienen al menos tres desventajas: 1) no tienen un significado físico claro y útil como las energías de un FB; 2) requieren una transformación lineal a partir de las energías de un FB o de los coeficientes LP; y 3) en los modelos ocultos de Markov (HMM)) con funciones de probabilidad gaussianas de matriz de covarianza diagonal, la forma de la ventana no tiene ningún efecto y sólo su longitud, es decir, el número de parámetros, es una variable de control.

Con el fin de superar estas inconvenientes, se ha propuesto recientemente una representación espectral alternativa que resulta de filtrar la secuencia frecuencial del logaritmo de las energías espectrales con un filtro FIR de orden 1 ó 2 tanto para reconocimiento del habla [6] como del locutor [7] [8]. En estos trabajos, se ha comprobado que el filtrado frecuencial puede ecualizar de forma aproximada la varianza cepstral, realzando las oscilaciones de la curva de la envolvente espectral que son más efectivas para la discriminación entre locutores.

El propósito de esta comunicación es mostrar el comportamiento de la hibridación de los análisis espectrales LP y FB (sección 2) y el filtrado frecuencial (sección 4) en reconocimiento del locutor. Se muestra que el filtrado frecuencial produce a la vez dos efectos deseados: descorrelación y discriminación (seccion 3). Aplicando filtrado frecuencial sobre el logaritmo de las energías de un FB, se han observado mejores resultados que usando mel-cepstrum en experimentos de identificación del locutor independiente del texto sobre las base de datos TIMIT, especialmente cuando se ha añadido ruido blanco (sección 5). También se han realizado experimentos de verificación dependientes del texto sobre la nueva base telefónica de locutores POLYCOST combinando filtrado frecuencial y análisis espectral híbrido y se han obtenido mejores resultados que utilizando las parametrizaciones convencionales LP-cepstrum y mel-cepstrum.

2. ANÁLISIS ESPECTRAL HÍBRIDO

El método LP está relacionado estrechamente con el modelo digital de producción del habla. De ahí que pueda esperarse de él un deconvolución adecuada entre la respuesta del tracto vocal y la excitación glotal.

Mientras que la LP no considera bandas en el espectro, el método FB elimina la información de tono y reduce la varianza de la estimación integrando el periodograma (cuadrado de la transformada discreta de Fourier) en bandas frecuenciales. El método FB modela separadamente la potencia en cada banda y ofrece la posibilidad de distribuir la posición de las bandas en el eje frecuencial (se utiliza una escala mel en el llamado mel-cepstrum) y definir su anchura y su forma como se desee para aprovechar las propiedades de percepción del sistema auditivo humano. Esta localización frecuencial de los parámetros da lugar a otras ventajas. Por ejemplo, si se conoce la relación señal a ruido de cada banda, puede usarse de forma directa: substracción espectral, enmascaramieto, etc.

La combinación de los dos análisis LP y FB puede proporcionar mejores parámetros espectrales. Una posible manera es aplicar análisis FB sobre la señal antes del análisis LP [9] [10]. Esta alternativa se denominará FB-LP y se implementará de forma similar a los coeficientes PLP [9], pero usando un orden de predicción mayor y sin ponderación perceptiva ni compresión de amplitud. Otra alternativa es aplicar LP seguida de un FB, a la cual se la denominará LP-FB.

Las parametrizaciones convencionales, LP-cepstrum y mel-cepstrum y las representaciones cepstrales de las dos técnicas híbridas FB-LP y LP-FB pueden incluirse en un esquema de parametrización unificado [2], que puede conducir a otras

nuevas técnicas de parametrización de la voz.

3. DESCORRELACIÓN Y DISCRIMINACIÓN

Los HMM suelen utilizarse con matrices de covarianza diagonales. En este caso, se asume implícitamente que los parámetros espectrales están incorrelados. Esto es cierto en modelos continuos y semicontinuos con funciones de densidad gaussianas y también para modelos discretos con distancia de Mahalanobis. Sin embargo, la secuencia frecuencial de los logaritmos de las energías en un FB S(k) está altamente correlada. El mel-cepstrum es una forma obtener a partir de S(k) un conjunto de parámetros casi incorrelados. De hecho, aproximando el proceso aleatorio S(k) por un modelo de Markov de primer orden, se concluye que la transformada discreta coseno es casi equivalente a la transformación de Karhunen-Loève.

La descorrelación es, pues, una propiedad deseada de los conjuntos de parámetros espectrales debido a la forma particular en que son usados en los sistemas actuales de reconocimiento. Desde luego, también debido a que puede proporcionar una representación menos redundante. Sin embargo, lo que es relevante realmente para el propio proceso de clasificación es la capacidad de discriminación de estos parámetros.

Es un hecho conocido que la varianza de C(m) decrece con m [4]. La Figura 1 muestra una estimación de esta varianza para la base de datos TIMIT utilizando Q=20 bandas frecuenciales mel. Notar el valor nulo correspondiente la cuefrencia cero, causado por la substracción del valor medio de S(k) [11].

Por tanto, las bajas cuefrencias m dominarán en general los cálculos de probabilidad o distancia en el clasificador. Podemos preguntarnos si puede mejorarse el reconocimiento mediante una ecualización adecuada de la varianza global

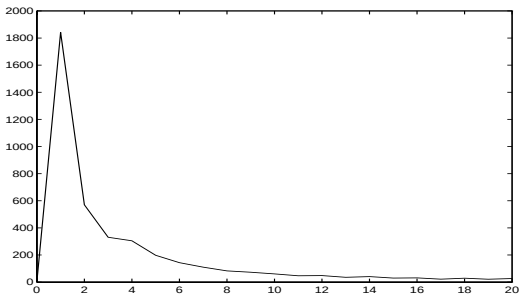

Figura 1.- Varianza cepstral para la base de datos TIMIT

de C(m), tal como ocurre en reconocimiento del habla [6]. Además, hay que destacar que existe una relación estrecha entre ecualización de la varianza de C(m) a bajas cuefrencias y descorrelación de S(k).

Sin embargo, una varianza plana puede no ser el objetivo más adecuado para el reconocimiento. Por ejemplo, cuando el intervalo frecuencial entre bandas no es suficientemente grande, esta ecualización pondera demasiado el ruido de estimación de C(m) a cuefrencias altas. Otra razón para no ecualizarla completamente puede ser la presencia de distorsiones de canal o ruido aditivo de banda ancha, que pueden requerir una atenuación mayor de las cuefrencias más bajas.

Una posible medida de la capacidad de discriminación de cada coeficiente cepstral puede ser el cociente entre su varianza inter-locutor y su varianza global. La Figura 2 muestra una estimación de este cociente para la base de datos TIMIT usando Q=20 bandas frecuenciales.

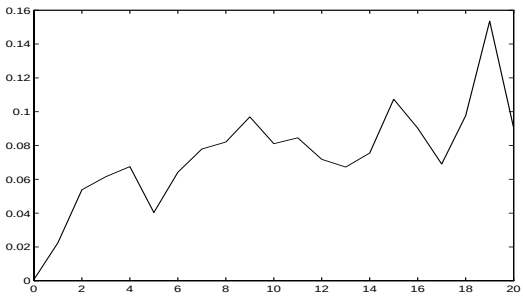

Figura 2.- Estimación del cociente entre las varianzas cepstrales inter-locutor y global para la base de datos TIMIT

Como puede verse en la Figura 2, el rango dinámico de la secuencia de este cociente es menor que el de la varianza global mostrada en la Figura 1. Este hecho sugiere que una ecualización aproximada de la varianza puede ayudar a incrementar la capacidad de discriminación de la secuencia cepstral, al menos para habla limpia.

Por otro lado, la Figura 2 muestra una ligera tendencia creciente con el índice cuefrencial m. Este hecho lleva a pensar que la información más discriminante está localizada en las cuefrencias más altas, es decir, en la alternancia rápida de picos y valles de la curva espectral y no en las cuefrencias más bajas. De hecho, la mayoría de los sistemas de reconocimiento del locutor utilizan un número mayor de parámetros cepstrales que los reconocedores del habla. Incluso podría ser conveniente enfatizar ligeramente las cuefrencias altas.

El liftado cepstral (ponderación en m) ha sido la forma usual de compensar el excesivo peso de los términos de m inferior tanto en reconocimiento del habla como del locutor. En este caso, se necesitan dos pasos para obtener los parámetros finales a partir de las energías de un FB: 1) una transformación lineal (transformada discreta coseno), que descorrela significativamente la secuencia de parámetros, y 2) una ponderación (liftado) de los coeficientes cepstrales. En los HMM con funciones de densidad gaussianas de matrices de covarianza diagonal, la forma de la ventana cepstral no tiene ningún efecto debido a la normalización intrínseca de varianza de la función gaussiana.

En trabajos recientes, para superar estos inconvenientes y para tener parámetros que posean un significado frencuencial, se introdujo una alternativa al uso de los parámetros cepstrales para reconocimiento del habla [6] y del locutor [7], que consiste en un procesado simple en el dominio del logaritmo de la energía. La transformación de la secuencia del logaritmo de las energías del FB a los parámetros cepstrales se evita realizando un filtrado de la misma, que en adelante se denominará filtrado frecuencial para denotar que la convolución se realiza en el dominio frecuencial.

Este filtrado frecuencial produce a la vez dos efectos, descorrelación y discriminación, utilizando un filtro FIR de primer o segundo orden. Además, el filtrado frecuencial puede ponderar el cepstrum de forma implícita en los HMM con funciones de densidad gaussianas de matrices de covarianza diagonales.

4. FILTRADO FRECUENCIAL

Se desea ecualizar de forma aproximada la varianza de los coeficientes cepstrales filtrando la secuencia frecuencial de los logaritmos de las energías. Implementando este filtrado como una convolución circular con la secuencia h(k), los coeficientes cepstrales quedarán así multiplicados por la DFT de h(k), que se denotará por H(m).

En primer lugar, como en el FB mel convencional no hay filtros centrados en $\omega=0$ y $\omega=\pi$, se añade un cero en los dos extremos de la secuencia, es decir, S(0)=S(Q+1)=0, para representar la poca energía contenida en estas bandas. Después, de acuerdo con la práctica común [11], en cada tramo de voz, se resta el valor medio de la secuencia asimétrica S(k), k=1,...., Q. Seguidamente, se convoluciona circularmente S(k) con h(k) para obtener la secuencia filtrada. Como en el sistema de reconocimiento sólo se usan los valores de la secuencia filtrada entre k=1 y k=Q, podemos emplear una h(k) de hasta longitud 2 en el cálculo de la segmento a utilizar de la secuencia filtrada sin que haya interferencia con las muestras de la secuencia simétrica S(k), k=1,..., Q. De este modo, podemos considerar este proceso como una filtrado lineal, con h(k) como respuesta impulsional.

Puede obtenerse fácilmente un filtro FIR de primer orden que ecualice al máximo la varianza de los coeficientes cepstrales utilizando el método de mínimos cuadrados

del siguiente modo. En primer lugar, se estima la varianza promediando sobre todos los tramos de la base de datos. Después, se aplica una DFT inversa para obtener la varianza de S(k) y se calcula el cociente r entre los valores en k=1 y 0 de la misma. El filtro FIR de primer orden que aplana máximamente la varianza será $H(z)= 1-rz^{-1}$.

La Figura 3 muestra, para la base de datos TIMIT con Q=20, el producto de la varianza cepstral global, mostrada en la Figura 1, por el módulo de la respuesta frecuencial muestreada H(m) obtenida de este modo. El valor de r resultante es 0.75. Como puede verse, se ha ecualizado de forma aproximada la tendencia de la varianza cepstral.

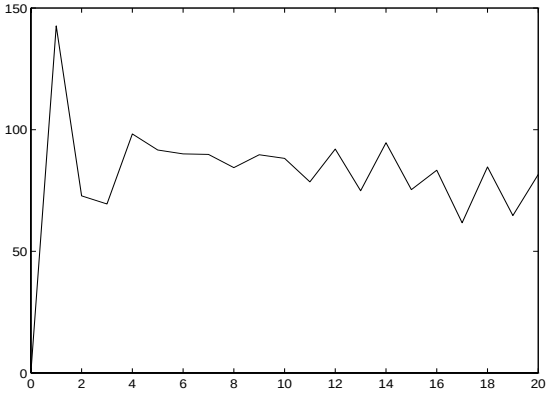

Figura 3.- Varianza cepstral ecualizada por la respuesta frecuencial muestreada

Sin embargo, el filtrado frecuencial puede mejorar sus prestaciones si se optimiza empíricamente el filtro, teniendo en cuenta por ejemplo la tendencia ligeramente creciente en m del cociente entre las varianzas cepstrales inter-locutor y global mostrado en la Figura 2.

En efecto, los parámetros espectrales que resultan del filtrado de la secuencia frecuencial del logaritmo de las energías de un FB han demostrado ser competitivas con respecto al mel-cepstrum [6] [7]. Sin embargo, el filtrado frecuencial no sólo puede aplicarse a las energías de un FB, sino también cuando se realiza un análisis LP, tal como se describió en [6] e incluso en el caso del análisis espectral híbrido descrito en la sección 2.

Finalmente, es importante destacar la simplicidad computacional del filtrado frecuencial con respecto al mel-cepstrum. Una forma de reducir aun más el coste computacional es utilizar los filtros $1-z^{-1}$ y $z-z^{-1}$, ya que no necesitan productos y hacen innecesaria la substracción del valor medio debido a su cero de transmisión en la frecuencia cero. El filtro de primer orden $1-z^{-1}$, que es equivalente a la ventana rampa [5], consiste simplemente en restar la muestra anterior a la banda actual. El filtro de segundo orden $z-z^{-1}$,que es equivalente a la ventana paso-banda [3], consiste en restar las dos muestras adyacentes a la actual. Estos filtros no dependen de la base de datos y proporcionan resultados similares a los del filtro óptimo, que depende de la base de datos [6].

5. EXPERIMENTOS DE IDENTIFICACIÓN

Se han llevado a cabo experimentos de identificación del locutor independiente del texto, tanto en ausencia como en presencia de ruido, utilizando como representación acústica el resultado del filtrado de la secuencia frecuencial del logaritmo de las energías de un FB, sin energía ni parámetros diferenciales adicionales.

Se utilizó la base de datos TIMIT. Para ello, se seleccionaron 200 locutores, 100 de cada sexo. Siempre se utilizó voz limpia para el entrenamiento. La voz ruidosa de prueba se simuló añadiendo ruido blanco gaussiano de media nula a la señal limpia de forma que la relación señal-ruido resultante fuera de 20 dB.

Se modificó el software HTK, basado en HMM continuos, para realizar experimentos de verificación del locutor con las nuevas representaciones acústicas. En la etapa de parametrización, después del preénfasis de la señal con un cero en z=0.95, se tomaron cada 10 ms tramos enventanados con una ventana de Hamming de 25 ms. Cada tramo se representó con M=20 parámetros derivados de un banco de Q=20 filtros. Se caracterizó cada locutor

con un HMM de 1 estado con 32 gaussianas de matriz de covarianza diagonal. El silencio se caracterizó con un HMM de 3 estados y 1 gaussiana. Para cada locutor, el modelo se entrenó con 5 frases. Las otras 5 se utilizaron separadamente como prueba.

La Tabla 1 muestra las tasas de identificación (ID) de locutor en condiciones limpias y ruidosas obtenidas con el mel-cepstrum convencional (MFCC) y el filtrado de logaritmo de las energías de un FB (FLFBE) usando varios filtros FIR paso-alto de primer orden: $1\text{-}0.75z^{-1}$, que ecualiza la base TIMIT para Q=20, tal como se muestra en las figuras anteriores; y $1\text{-}0.8z^{-1}$, $1\text{-}0.9z^{-}1$ y $1\text{-}z^{-1}$, inspirados en la tendencia creciente de la curva del cociente entre las varianzas cepstrales inter-locutor y global de la Figura 2.

Parameters / ID	clean	20 dB
MFCC	98.1	32.4
FLFBE ($1\text{-}0.75z^{-1}$)	98.3	46.1
FLFBE ($1\text{-}0.8z^{-1}$)	98.5	52.8
FLFBE ($1\text{-}0.9z^{-1}$)	98.4	61.8
FLFBE ($1\text{-}z^{-1}$)	98.3	64.4

Tabla 1.- Tasas de identificación

Puede verse en la Tabla 1 que la nueva parametrización FLFBE es competitiva con respecto a la representación convencional MFCC en condiciones limpias de ruido. Cuando se utiliza el ecualizador óptimo para la base de datos TIMIT, FLFBE supera al mel-cepstrum. Sin embargo, los mejores resultados se obtienen usando un filtro, que enfatiza ligeramente las cuefrencias mayores con respecto al cepstrum ecualizado.

El filtro $z\text{-}z^{-1}$, que proporcionó resultados cercanos al filtro óptimo en reconocimiento de habla limpia [6], no ha proporcionado tan buenos resultados en nuestra tarea: una tasa de identificación del 97.8 %. Ello puede deberse a las características paso-banda de este filtro. Los filtros paso-alto, como los considerados en la Tabla 1, son los adecuados si se quiere enfatizar las cuefrencias altas.

En condiciones ruidosas se han obtenido excelentes resultados usando el nuevo método FLFBE. Con el ecualizador óptimo $1\text{-}0.75z^{-1}$, hay una reducción de la tasa de error de identificación de casi un 30% respecto al mel-cepstrum. Los resultados son todavía mejores usando filtros que ponen más énfasis en las cuefrencias altas. Poniendo el cero en z=1 (filtro $1\text{-}z^{-1}$), hay una reducción de la tasa de error de casi un 50 %. Los filtros con cero cerca de 1 son mejores en presencia de ruido de banda ancha porque los parámetros cepstrales de menor índice son más afectados en general por este tipo de ruido que los de índice mayor.

6. EXPERIMENTOS DE VERIFICACIÓN

Para complementar estos experimentos, se han realizado experimentos de verificación del locutor dependientes del texto combinando filtrado frecuencial y análisis espectral híbrido sobre la nueva base de datos POLYCOST,

La base POLYCOST ha sido grabada sobre línea telefónica y diseñada para tareas de reconocimiento de locutores. Ha sido una iniciativa común de la acción europea COST 250, de nombre "Reconocimiento del locutor en telefonía". Se grabó a través de la red telefónica europea entre enero y marzo de 1996. Se utilizó una frecuencia de muestreo de 8 kHz. La base contiene unas 10 sesiones de 134 locutores de 14 países. Cada sesión consta de 15 elocuciones: una para detección DTMF, 10 de dígitos conectados en inglés, 2 con frases fijas en inglés y 2 libres en lengua materna. La lengua materna de la mayoría de locutores no es el inglés, lo cual posibilita experimentos sobre variabilidad intra/inter-locutor/lengua.

Se eligió un experimento de verificación sobe una frase fija, que fue común para todos los locutores, concretamente "*Joe took father's green shoe bench out*".

Se construyó un modelo de cliente por locutor a partir de las 4 primeras sesiones y un modelo global a partir de las 5 primeras sesiones de 22 locutores que se habían apartado previamente de la base. Las pruebas de clientes se hicieron con la 5ª sesión y las siguientes. Con las sesiones existentes para los 110 clientes elegidos, se pudieron realizar 660 pruebas de clientes. Para simular los intentos de impostores contra el locutor X, se utilizó la 5ª sesión de todos los locutores de la base excepto el X. Con 109 pruebas de impostor por cliente, se realizaron 11990 pruebas de impostor.

En la etapa de comparación se utilizó un software desarrollado en el proyecto CAVE, que se basa en los métodos descritos en el libro EAGLES [12].

Se usó también el sistema de reconocimiento HTK con la misma topología que en los experimentos de identificación. En la etapa de parametrización, la señal de voz (sin preénfasis) se fragmentó en tramos de 20 ms a un ritmo de 10ms, y cada tramo se caracterizó con M=20 parámetros obtenidos por una de las técnicas de análisis espectral consideradas anteriormente -LP, FB, FB-LP, LP-FB- y usando transformación cepstral o filtrado frecuencial. Cuando se aplicó análisis LP, se fijó el orden de predicción a 20. Cuando se usó un FB se utilizaron 2 filtros. No se utilizó ni energía ni parámetros delta.

La Tabla 2 muestra los resultados de verificación de locutor en términos de tasa de igual error (EER). Se consideraron varios filtros FIR paso-alto de primer orden en el caso del filtrado frecuencial: $1\text{-}0.5z^{-1}$, que ecualiza la varianza del mel-cepstrum en los dígitos aislados de la partición de adultos de la base TI para M=Q=12, tal como se usó en [6]; $1\text{-}0.75z^{-1}$, que ecualiza la varianza del mel-cepstrum en la base de datos TIMIT para M=Q=20; y $1\text{-}z^{-1}$, inspirado en la tendencia creciente de la curva de la Figura 3. También se probó el filtro paso-banda de segundo orden $z\text{-}z^{-1}$. Los resultados de verificación de la Tabla 2 se han representado gráficamente en la Figura 4.

Anal.	C	$1\text{-}rz^{-1}$ r=0.5	$1\text{-}rz^{-1}$ r=0.75	$1\text{-}z^{-1}$	$z\text{-}z^{-1}$
FB	3.75	4.22	4.67	4.88	2.55
LP	3.40	3.10	2.92	2.65	2.96
FB-LP	3.48	3.04	2.68	3.05	2.71
LP-FB	3.41	4.44	4.14	4.63	3.46

Tabla 2.- Tasas de verificación en términos de EER

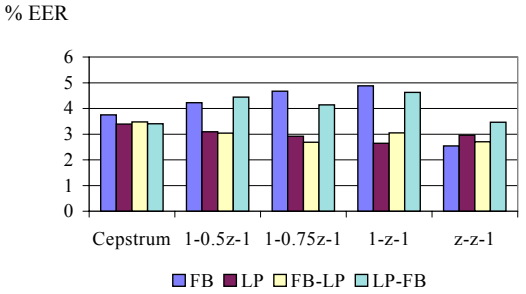

Figura 4.- Representación gráfica de los tasas de verificación

Como puede verse, cuando se utiliza la transformación cepstral, los mejores resultados se obtienen usando LP. El LP-cepstrum convencional obtiene 3.40% EER, mientras que el mel-cepstrum proporciona un 3.75%. Los análisis espectrales híbridos proporcionan resultados intermedios: FB-LP-cepstrum proporciona 3.48% y LP-FB 3.41%.

Con respecto al uso de la nueva técnica de filltrado frecuencial, los resultados dependen drásticamente del tipo análisis espectral. En el caso de análisis FB, el único filtro que supera la transformación cepstral el filtro paso-banda de segundo orden $z\text{-}z^{-1}$. Este resultado no es acorde con las conclusiones previas acerca de la conveniencia de filtros paso-alto. Sin embargo, usando este filtro paso-banda se consigue un 2.55%, los mejores resultados obtenidos en este trabajo. La mejora

relativa con respecto al mel-cepstrum convencional es de un 32%.

En el caso de análisis LP, todos los filtros paso-alto de primer orden superan a la transformación cepstral. El mejor resultado, un 2.65% EER, se obtiene usando $1-z^{-1}$. En este caso, la mejora relativa con respecto al LP-cepstrum es de un 22%.

Con respecto al análisis híbrido espectral, el comportamiento del filtrado frecuencial es bastante distinto. En el caso de análisis LP-FB, el uso del filtrado frecuencial no mejora la transformación cepstral. En cuanto al análisis FB-LP, se obtiene un 2.71% EER usando el filtro paso-banda $z-z^{-1}$ y un 2.69% EER usando el filtro paso alto $1-0.75z^{-1}$. En este caso, la mejora relativa con respecto la FB-LP-cepstrum es de un 23%.

7. CONCLUSIONES

En esta comunicación, se han probado dos formas de obtener parámetros acústicos robustos para el reconocimiento automático de locutores: la hibridación de la predicción lineal (LP) y el banco de filtros (FB), y el filtrado frecuencial del logaritmo de las energías espectrales como alternativa al cepstrum. Esta combinación, que había superado a las técnicas convencionales en reconocimiento de habla limpia y ruidosa, ha proporcionado buenas tasas de identificación y verificación de locutores. Los mejores resultados de verificación en la base telefónica de locutores POLYCOST se han obtenido usando un filtro paso-banda de segundo orden para el análisis espectral FB y un filtro paso-alto de primer orden para los análisis LP y FB-LP (FB antes de LP).

REFERENCIAS

[1] J. Thompson, J.S. Mason, "Within Class Optimization of Cepstra for Speaker Recogniton", Proc. EUROSPEECH'95, pp. 165-168, Madrid, 1995.

[2] J. Hernando, C. Nadeu,, "Robust Speech Parameters Located in the Frequency Domain", Proc. EUROSPEECH'97, pp. 417-420,Rodas (Grecia), 1997

[3] B. H. Juang, L.R. Rabiner, J. G. Wilpon, "On the Use of Banspass Liftering in Speech Recognition", IEEE Trans. ASSP, vol. 35, nº 7, pp. 947-953.

[4] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition", IEEE Trans, ASSP, vol. 35, nº 10, Octubre 1987.

[5] B.A. Hanson, H. Wakita, "Spectral Slope Based Distortion Measures for All-Pole Models of Speech", IEEE Trans. ASSP, vol. 35, nº 7, pp. 968-973, 1987.

[6] C. Nadeu, J. Hernando, M. Gorricho, "On the Decorrelation of Filter-Bank Energies in Speech Recognition", Proc. EUROSPEECH'95, pp. 1381-1384, Madrid, 1995.

[7] J. Hernando, C. Nadeu, "CDHMM Speaker Recognition by means of Frequency Filtering of Filter-Bank Energies", Proc. EUROSPEECH'97, pp. 2363-2366, Rodas (Grecia), 1997.

[8] J. Hernando, C. Nadeu, "Speaker Verification on the POLYCOST database using frequency filtered spectral energies, Proc. ICSLP'98, pp. 129-132, Sidney (Australia), 1998.

[9] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", JASA, Vol. 87, No. 4, pp. 1738-52, 1990.

[10] M.G. Rahim, B.H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", IEEE Trans. SAP, Vol. 4, No. 1, pp. 19-30, 1996.

[11] J.W. Picone, "Signal Modeling Techniques in Speech Recognition", Proc. IEEE, Vol.81, No.9, Sept.1993, pp. 1215-47

[12] F. Bimbot, G. Chollet, "Assesment of Speaker Verification Systems", en *Spoken Resources and Assessment, EAGLES Handbook.*