TECNIACUSTICA
BARCELONA • 1996

Técnicas robustas para la discriminación de locutores

*Jordi Muñoz, Javier Hernando, Climent Nadeu, Pau Pachès**

Dep. Teoría de la Señal y Comunicaciones. Universidad Politécnica de Catalunya
javier@gps.tsc.upc.es

ABSTRACT
Recently, a new filtering technique based on the decorrelation of filter bank energies has shown to be attractive for speech recognition because of its simplicity and its lower computational cost than standard representations like LPC-cepstrum or mel-cepstrum. Experimental results obtained in speaker identification show that this filtering technique is useful for discriminating speakers. On the other hand, different normalization techniques are applied and compared in speaker verification, and experimental results are shown.

INTRODUCCION

Los sistemas de reconocimiento de locutor son de interés creciente por sus variados usos, como son aplicaciones periciales para la identificación de locutor o aplicaciones de seguridad para la verificación de locutor. A parte de la clasificación anterior segn la función que realizan, también pueden dividirse según el tipo de mensaje utilizado en sistemas dependientes del texto o independientes del texto. Por último se puede hacer una división entre sistemas cerrados, cuando todos los locutores que tienen acceso al sistema son conocidos por él, o abiertos, cuando no ocurre lo anterior.

Mejorar el funcionamiento de estos sistemas implica aumentar la discriminación entre locutores, lo cual es lo mismo que enfatizar la varianza de las características acústicas propias del locutor entre los diferentes locutores (información de locutor), frente a la varianza de las características acústicas pertenecientes a la información semántica (información del mensaje). Para realizar lo anterior se han realizado pruebas de identificación de locutor utilizando el filtrado frecuencial del logaritmo de las energías de un banco de filtros [1], de uso reciente y con menor coste computacional que otras parametrizaciones estándar Por ltimo se exploran las técnicas de normalización aplicadas en la verificación de locutor como una herramienta útil para la discriminación en estos sistemas.

En el apartado 2 se introducen los fundamentos teóricos del filtrado frecuencial. Las técnicas de normalizacin para la verificación se exponen en el apartado 3. El apartado 4 describe las pruebas realizadas mostrando los resultados de identificación con el uso del filtrado frecuencial, y resultados de verificación de los distintos tipos de normalización. Por último, en el apartado 5 se exponen las principales conclusiones obtenidas y en el apartado 6 se citan las referencias bibliográficas.

FILTRADO FRECUENCIAL

Es bien conocido que los coeficientes cepstrales c[m], obtenidos a partir del análisis LPC (LPC-cepstrum) o del logaritmo de las energías de un banco de filtros en escala mel (mel-cepstrum), son de una amplia utilización en sistemas de procesado de voz debido a su alto grado de incorrelación (no presentan redundancia en su información) y a la fiable descripción que aportan del tracto vocal debido a la deconvolución homomórfica. Sin embargo poseen una varianza que decae con el índice del coeficiente m [1], lo cual produce que los coeficientes de menor índice tengan más peso en el reconocimiento. Para evitar esto es muy frecuente el uso del enventanado de los coeficientes cepstrales c[m], que se conoce como *liftering*, y con el cual se consigue compensar el efecto anterior.

---

[1] Este trabajo ha sido financiado por los proyectos TIC95-0884-C04-02 y TIC 95-1022-C05-03

Cuando se utiliza como herramienta clasificadora a los Modelos Ocultos de Markov o a los Modelos de Mezcla de Gaussianas, si las funciones de densidad de probabilidad de observación se construyen con matrices de covarianza diagonales, el *liftering* no tiene ningún efecto. Si convolucionamos el logaritmo de las energías de un banco de Q filtros (log FBE), $S(k)$, con un filtro $h_f(k)$, es equivalente a un enventanado en el dominio del cepstrum [1]

$$S(k) \otimes h_f(k) \overset{DFT^{-1}}{\rightarrow} c[m] \cdot H_f[m] \quad (1)$$

donde $H_f[m]$ es la *DFT* inversa de $h_f(k)$, actuando como ventana ponderadora de los coeficientes cepstrales, consiguiéndose con este procedimiento el efecto ponderador sin que pueda ser eliminado por las funciones de densidad de probabilidad de observación gaussianas con matrices de covarianza diagonales. Antes de efectuar el filtrado frecuencial se sustrae la media del logaritmo de las energías del banco de filtros, lo que equivale a eliminar el contenido energético de la señal, haciendo a la parametrización independiente del volumen de la voz del locutor. Esto es muy importante en el reconocimiento de voz y también en el reconocimiento de locutor debido a la variabilidad de los niveles de energía de la señal por parte del mismo locutor. El filtro frecuencial que ecualiza la varianza de los coeficientes cepstrales produce una decorrelación del logaritmo de las energías del banco de filtros haciendo a esta parametrización muy útil para el clasificador [1]. La utilización del filtrado frecuencial presenta un menor coste computacional que la parametrización de coeficientes cepstrales LPC-cepstrum o mel-cepstrum.

NORMALIZACION EN LA VERIFICACION

El uso de la normalización de las probabilidades acumuladas en la verificación es de uso extendido por la reducción del efecto indeseado de la variabilidad del locutor a lo largo del tiempo. Se calcula si la probabilidad de que la señal observada **O** sobre el modelo del locutor que supuestamente está accediendo al sistema (locutor cliente C), normalizada por la probabilidad de un modelo que debe contemplar a la población impostora ($\overline{C}$), es mayor o no que un cierto umbral (para decidir si se acepta al locutor o se rechaza)

$$\log \hat{P}(O \mid \lambda_C) - \log \hat{P}(O \mid \lambda_{\overline{C}}) > \zeta \quad (2)$$

donde se usan logaritmos de probabilidades normalizados (penalizados) por la duración de la observación en tramas. Pueden adoptarse varias estrategias para construir el modelo de la población impostora. La primera que adoptaremos es la construccin de un modelo independiente del locutor (mil) a partir de señal de varios locutores. La segunda será el asignar agrupaciones de locutores cercanos al cliente, agrupaciones de locutores lejanos al cliente y combinaciones de las dos anteriores (agrupaciones mixtas), y obtener una probabilidad que represente a la población impostora a partir de la combinación de las probabilidades de observación de la señal sobre cada uno de los modelos de la agrupación. Las técnicas utilizadas para la construcción de las agrupaciones serán *maximally spread close* para las cercanas y *maximally spread far* para las lejanas según [2], con lo que se consigue una mejor estimación de la población impostora. Adoptaremos la estimación de la probabilidad sobre la población impostora para una agrupación de B locutores de dos maneras. La primera es la aproximación por la media aritmética utilizada para agrupaciones mixtas en [2] y que consiste en calcular la probabilidad de normalización como la media aritmética de las probabilidades sobre cada miembro de la agrupación. La segunda es la aproximación por la media geométrica que es la que mejor resultados obtuvo para agrupaciones cercanas en [3].

RESULTADOS EXPERIMENTALES

Se ha implementado dos sistemas de simulación tanto para identificación como para verificación de locutor, uno dependiente del texto trabajando sobre la base de datos de dígitos conectados en inglés de Texas Instruments, y otro independiente del texto que opera sobre la base de datos de habla continua en inglés TIMIT. El clasificador utilizado está basado en los Modelos Ocultos de Markov (HMM) y ha sido implementado mediante el software HTK [4] con las modificaciones pertinentes para la construcción de los sistemas. Se adoptó un periodo de trama de 10 ms, con ventana Hamming de 25 ms y coeficiente de preénfasis 0.95. Para el sistema dependiente del texto se ha creado un modelo izquierda-derecha de 5 estados para cada dígito de cada locutor, más otro modelo independiente del locutor para el silencio de 3 estados. Se utilizaron en el entrenamiento 10 elocuciones de 5 dígitos y 5 de 7 dígitos, y para el test 5 de 7 dígitos. La duración aproximada de cada elocución es de 2,4 s para las de 5 dígitos y de 2,7 s para las de 7 dígitos. Además se utilizó un modelo de cada dígito independiente del locutor para la normalización construido con toda la señal de entrenamiento. Para el sistema independiente se modeló a cada locutor con un modelo de un estado y 32 mezclas, un modelo para el silencio izquierda-derecha de 3 estados y un modelo independiente del locutor para la normalización de 1 estado y 64 mezclas. Utilizándose 5 elocuciones para el entrenamiento y 5 para el test, siendo la duración aproximada de cada elocución de 3 s.

Se realizaron pruebas de identificación de 100 locutores, con vectores de parámetros de dimensión 12, aplicando el filtro frecuencial FIR de primer orden siguiente

$$H_f(z) = 1 - a \cdot z^{-1} \quad (3)$$

siendo un parámetro a ajustar. Los mejores resultados obtenidos para el filtrado frecuencial (log FBE+ff), con $a = 5$ para el sistema dependiente del texto y $a = 6$ para el sistema independiente del texto, se muestran en la Tabla I junto a las parametrizaciones LPC-cepstrum, mel-cepstrum y para el logaritmo de las energías de un banco de filtros sin sustracción de la media ni filtrado frecuencial (log FBE).

Tabla I

	LPC-cepstrum	mel-cepstrum	log FBE	log FBE + ff
Dependiente texto	99,60	98,00	82,00	98,20
Independiente texto	98,60	98,00	9,80	98,80

Para comparar las técnicas de normalización de probabilidades sobre la parametrización LPC-cepstrum de 24 coeficientes se realizaron pruebas de verificación de 100 locutores clientes. Para el test se ha utilizado dos elocuciones por cada locutor lo que hace un total de 200 accesos de clientes, mientras que los impostores son los mismos 100 locutores que se hacen pasar por cada uno de los restantes locutores, exceptuando aquellos para los que el locutor que se hace pasar por impostor forma parte de alguna de sus dos agrupaciones. Esto hace un total de 7900 accesos de impostores. Después de realizar el test se ajusta el umbral para obtener la tasa de *equal error* (tasa de falsa aceptación igual a la tasa de falso rechazo) para evaluar el sistema de verificación. Los resultados de las tasas de igual error se muestran en la Tabla II para los distintos tipos de normalización y agrupaciones de 6 locutores.

Tabla II

	sin normalizar	mil	6 cercanos		6 lejanos		mixto 3 cercanos y 3 lejanos	
			aritmética	geométrica	aritmética	geométrica	aritmética	geométrica
Dependiente texto	2,91	2,09	1,00	1,52	1,82	1,82	1,00	1,00
Independiente texto	4,00	3,50	2,50	3,50	4,21	4,95	2,50	2,87

Las tasas de *equal error* de verificación en función del número de locutores en la agrupación para el sistema dependiente del texto se muestran en la figura 1, y para el sistema independiente del texto en la figura 2.

CONCLUSIONES

El filtrado frecuencial presenta un excelente comportamiento en el sistema independiente del texto mejorando a las demás parametrizaciones utilizadas. Para el sistema dependiente del texto no se consigue superar a la parametrización LPC-cepstrum, la cual ofrece un resultado muy elevado, pero sí a la mel-cepstrum. Se ha de considerar el importante ahorro computacional que supone la no necesidad de realizar la transformación lineal de paso al dominio cepstral.

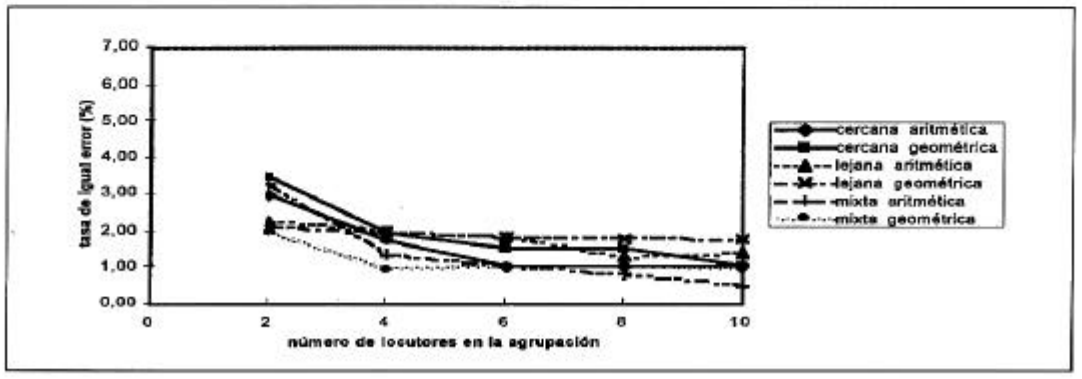

Figura 1

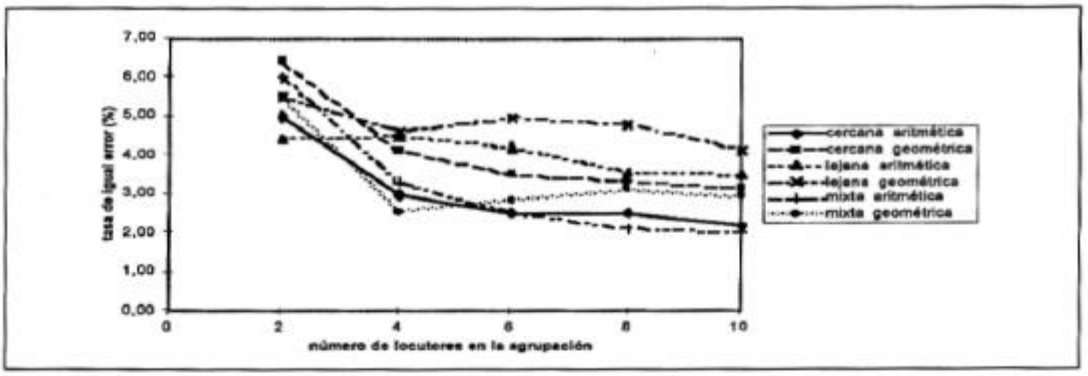

Figura 2

La normalización con agrupaciones de locutores se comporta mejor que el modelo independiente del locutor, las agrupaciones mixtas son las que mejor se comportan, siendo la aproximación geométrica la mejor para agrupaciones de pocos locutores (< 6), y la aritmética la mejor para agrupaciones mayores. Las agrupaciones cercanas con aproximación aritmética no difieren mucho de las mixtas y el uso de agrupaciones lejanas es desaconsejable.

REFERENCIAS

[1] C. Nadeu, J. Hernando, M. Gorricho, *"On the Decorrelation Of Filter-Bank Energies in Speech Recognition"*, Proc. EUROSPEECH'95, pp. 1381-4.

[2] D. A. Reynolds, *"Speaker identification and verification using Gaussian mixture speaker models"*, Speech Communication, vol. 17, 1995, pp. 91-108.

[3] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, F. K. Soong, *"The Use of Cohort Normalized Scores for Speaker Verification"*, International Conference Speech and Language Processing, November, 1992, pp. 599-602.

[4] Cambridge University Engineering Department, Speech Group and Entropic Research laboratories Inc., HTK: Hidden Markov Model Toolkit V1.5, December 7, 1993.