UNIVERSIDAD POLITECNICA DE CATALUÑA

Departamento de Teoria de la señal y comunicaciones

TECNICAS DE PROCESADO Y REPRESENTACION DE LA SEÑAL DE VOZ PARA EL RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS

Autor: Francisco Javier Hernando Pericas
Director: Climent Nadeu i Camprubi

Barcelona, mayo 1993

Capítulo 4

MEDIDAS DE DISTANCIA ROBUSTAS

En la etapa de parametrización de un sistema de reconocimiento automático del habla, tal como se ha descrito en el apartado anterior, la señal de voz se representa mediante una sucesión de vectores de parámetros acústicos con la información suficiente para poder identificar los sonidos en las siguientes etapas del sistema de reconocimiento.

En la inmensa mayoría de los sistemas de reconocimiento es necesario definir explícitamente una medida de distancia o distorsión entre estos vectores de parámetros para llevar a cabo la comparación de las señales de test y de referencia e identificar el mensaje a reconocer. Así, por ejemplo, en la aproximación clásica de comparación de patrones es necesaria una medida de distancia en el algoritmo de programación dinámica. Los modelos ocultos de Markov discretos y de múltiple etiquetado que han sido utilizados en las pruebas experimentales realizadas en este trabajo requieren una etapa de cuantificación vectorial, que exige también la definición de una medida de distancia entre vectores (ver apartado 5.2.3.2 de esta memoria).

Existe una estrecha relación entre la técnica de análisis espectral utilizada en la etapa de parametrización y la medida de distancia que se ha de aplicar para comparar los vectores resultantes de esta técnica concreta de parametrización. Es evidente que la distancia empleada ha de ser consistente con las características que la técnica de

parametrización confiere a estos vectores y la elección de una medida de distorsión adaptada a la parametrización es vital para obtener buenas tasas de reconocimiento.

Ello ha motivado el estudio en esta tesis de las medidas de distancia robustas adaptadas a las técnicas de parametrización presentadas en el capítulo anterior. El objetivo será encontrar medidas de distancia que reflejen exactamente las diferencias fonéticas entre espectros de voz, sean robustas por sí mismas al ruido y computacionalmente eficientes.

Como se ha visto en el capítulo anterior, la predicción lineal de la señal de voz es la técnica de parametrización más utilizada en la actualidad debido a su correspondencia con el modelo de producción de la señal de voz y a su eficiencia y prestaciones. Se han propuesto un gran número de distancias adaptadas al modelo de predicción lineal: de Itakura-Saito [Ita68], *Log Likelihood Ratio* o de Itakura [Ita75], de coseno hiperbólico [Gra76], etc.

Para aumentar la robustez frente al ruido de estas distancias ha sido considerada ampliamente la ponderación de las regiones de picos espectrales, que están menos afectados por el ruido. Entre este tipo de distancias destacan las *Weighted Likelihood Ratio* [Shi82], *Low Frequency Weighted Likelihood Ratio* [Sug82], *Unsymetrical Weighted Likelihood Ratio* [Mat86] y *Frequency-Weigted Itakura* [Soo87].

Sin embargo, el tipo de distancia más empleada por su eficiencia y buen comportamiento es el de las medidas de distancias euclídeas cepstrales ponderadas. Cuando se utilizan ventanas de ponderación adecuadas, estas distancias han resultado ser ventajosas para el reconocimiento del habla tanto en condiciones limpias como ruidosas [Han86] [Jua87a].

Como resultado de la ponderación cepstral se obtiene una versión suavizada del espectro que depende tanto de la forma y longitud de la ventana de ponderación como del orden del modelo de estimación espectral. Uno de los objetivos de este trabajo será, por tanto, obtener el grado óptimo de suavizado en condiciones ruidosas para el caso de la predicción lineal clásica y las parametrizaciones alternativas propuestas.

Por otro lado, evidencias tanto analíticas como experimentales indican que el ruido blanco aditivo provoca una reducción de la norma del vector cepstral utilizado en reconocimiento (término del origen excluido) pero deja la orientación del vector más o menos intacta [Man89b]. La reducción de la norma del vector es perjudicial cuando se

utiliza la distancia euclídea cepstral. Estos resultados sugieren el uso de la operación proyección entre vectores cepstrales para formular varias medidas de distorsión robustas. En este trabajo se considera el estudio de las distancias de proyección cepstral, utilizando distintos tipos de ponderación y órdenes del modelo de estimación espectral.

Además, se ha observado que las condiciones adversas afectan más a las representaciones espectrales instantáneas de la señal de voz que a las representaciones dinámicas. Por ello, paralelamente al estudio de nuevas representaciones y distancias espectrales, otro objetivo de este trabajo es profundizar en la implementación de los parámetros regresivos: algoritmo e intervalo de estimación, número óptimo de parámetros,... y se ha considerado la posibilidad de una generalización de esta técnica hacia un filtrado cepstral. Por último, también se ha estudiado la incorporación de la información de la energía de la señal de voz para robustecer el sistema de reconocimiento, tanto mediante representaciones instantáneas como dinámicas.

La estructura de este capítulo es la siguiente. En el apartado 4.1. se revisa el concepto de cepstrum y la obtención de los coeficientes cepstrales asociados al modelo de predicción lineal. En apartado 4.2, a partir de las características estadísticas de estos coeficientes, se deriva la distancia de Mahalanobis, cuya simplificación y generalización da lugar a las distancias euclídeas cepstrales ponderadas, descritas en el apartado 4.3. El apartado 4.4 está dedicado a las distancias de proyección cepstral. Finalmente, el apartado 4.5 aborda el tema de la utilización de la energía y las características dinámicas de la señal de voz.

4.1. CEPSTRUM

El cepstrum c(n) de la señal de voz se define como la transformada inversa de Fourier del logaritmo de su espectro localizado S(ω), es decir,

$$c(n) = F^{-1} \{\ln S(\omega)\} \qquad (4.1)$$

El término cepstrum es indicativo de haber realizado una transformación inversa del *spectrum* (espectro). La variable independiente del cepstrum se denomina cuefrencia, término formado a partir de la palabra frecuencia, y tiene carácter temporal. La

principal característica del cepstrum es que permite separar del espectro de la señal de voz la estructura fina y los formantes.

A partir del modelo lineal de producción del habla, se ha visto en el capítulo anterior que

$$S(\omega) = S_{uu}(\omega)\ |H(e^{j\omega})|^2, \tag{4.2}$$

donde $S_{uu}(\omega)$ es el espectro de la excitación u(n) y $H(e^{j\omega})$ es la respuesta frecuencial del filtro del tracto vocal. Por tanto, se puede escribir

$$\ln S(\omega) = \ln S_{uu}(\omega) + \ln |H(e^{j\omega})|^2. \tag{4.3}$$

y la expresión del cepstrum será

$$c(n) = F^{-1}\ \{\ln S_{uu}(\omega)\} + F^{-1}\ \{\ln |H(e^{j\omega})|^2\}. \tag{4.4}$$

El primer término corresponde a la estructura fina del espectro. En el caso de voz sonora presenta un pico en la región de altas cuefrencias y a partir de él puede detectarse el tono de la voz. También se ha propuesto la extracción del tono a partir del cepstrum de la parte causal de la autocorrelación de la señal [Nad91]

El segundo término corresponde a la envolvente espectral y, por ello, se concentra en la región de bajas cuefrencias, de 0 a 2 ó 4 ms. La transformada de Fourier de las componentes de baja cuefrencia es, por tanto, el logaritmo de la envolvente espectral. El índice máximo de las componentes de baja cuefrencia usados en la transformada determina la suavidad de la envolvente espectral. Este método de estimar la envolvente del espectro se conoce como suavizado cepstral.

El proceso de separar las componentes cepstrales en estos dos factores se denomina liftado (*liftering* en inglés, derivado de *filtering*, filtrado) y consiste sencillamente en un enventanado. El análisis cepstral permite de este modo convertir la ecuación de convolución (3.5) en un suma. Es, pues, un caso particular de procesado homomórfico [Opp75].

En la figura 4.1 está representada a) una trama de voz sonora, b) su cespstrum obtenido utilizando FFT´s y c) las componentes cepstrales de alta cuefrencia ampliadas,

donde se puede observar el tono de la señal de voz M_0. Las componentes cepstrales de baja cuefrencia, que han sido suprimidas en la gráfica c) corresponden a la envolvente del espectro de la señal.

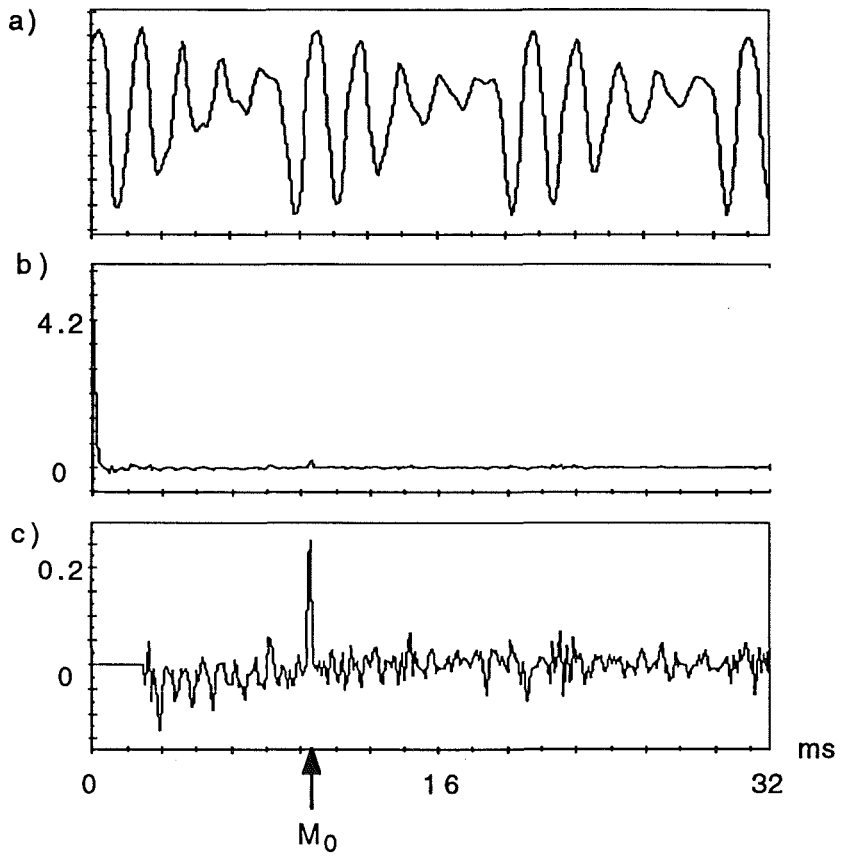

Fig. 4.1. Cepstrum de la señal de voz: a) trama de voz sonora, b) cepstrum, c) componentes de alta cuefrencia ampliadas.

Si se parte de un modelo LPC estable de la señal de voz

$$S(\omega) = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-j\omega k}\right|^2}, \qquad (4.5)$$

es fácil comprobar que los coeficientes cepstrales responden a la siguiente recursión

$$c(n) = c(-n) \qquad n < 0 \qquad (4.6)$$

$$c(0) = \ln (G^2) \qquad (4.7)$$

$$c(1) = - a_1 \qquad (4.8)$$

$$c(n) = - a_n - \sum_{m=1}^{n-1} \left(1 - \frac{m}{n}\right) a_m c(n-m) \qquad n = 1,\ldots p \qquad (4.9)$$

$$c(n) = - \sum_{m=1}^{p} \left(1 - \frac{m}{n}\right) a_m c(n-m) \qquad n > p. \qquad (4.10)$$

Este cepstrum es conocido como cepstrum LPC. La envolvente espectral que se obtiene liftando y transformando este cepstrum es una versión suavizada del espectro LPC, que ya es en sí mismo una estimación de la envolvente de la señal de voz. Debido a que el modelo LPC pondera de forma especial los picos del espectro, esta envolvente tiende a seguir los picos de forma más estricta que la envolvente espectral obtenida a partir del cepstrum calculado directamente a partir de la señal utilizando FFT, que suele conocerse como cepstrum FFT [Fur89].

4.2. DISTANCIA DE MAHALANOBIS

Debido a que el cepstrum es una secuencia par (4.6) y el valor en el origen es el logaritmo de la ganancia del modelo (4.7), en reconocimiento del habla sólo se utilizan los coeficientes cepstrales a partir de n=1. Además, debido a que el espectro LPC conserva únicamente la envolvente de la señal de voz, la magnitud de los coeficientes cepstrales LPC decae fuertemente a partir del origen y el suavizado cepstral del espectro LPC no requiere un excesivo número de términos antes de que el espectro suavizado aproxime al modelo [Gra76]. Por ello, el vector de parámetros acústicos o vector de características usado en reconocimiento es un vector de longitud L formado por los coeficientes cepstrales c(n), de n=1 a L, donde el valor de L es optimizado en función de los resultados de reconocimiento.

Las distribuciones estadísticas de los coeficientes cepstrales, estimadas para un conjunto amplio de locutores, muestran un comportamiento de tipo gaussiano. En la figura 4.2 se muestran los histogramas de los 10 primeros coeficientes cepstrales LPC, de c(1) a c(10), obtenidos a partir de las pronunciaciones de 10 locutores de un vocabulario de 117 palabras [Gom90].

Se puede plantear el problema de la definición de la distancia entre vectores de características desde el punto de vista probabilístico. Cada clase de vectores se

corresponde con un determinado sonido. Los vectores de características de cada clase tienen en este caso funciones de densidad de probabilidad gaussianas multivariadas definidas sobre un espacio vectorial multidimensional. En la práctica, las funciones asociadas a diferentes clases se solapan, lo cual es una fuente de errores de reconocimiento.

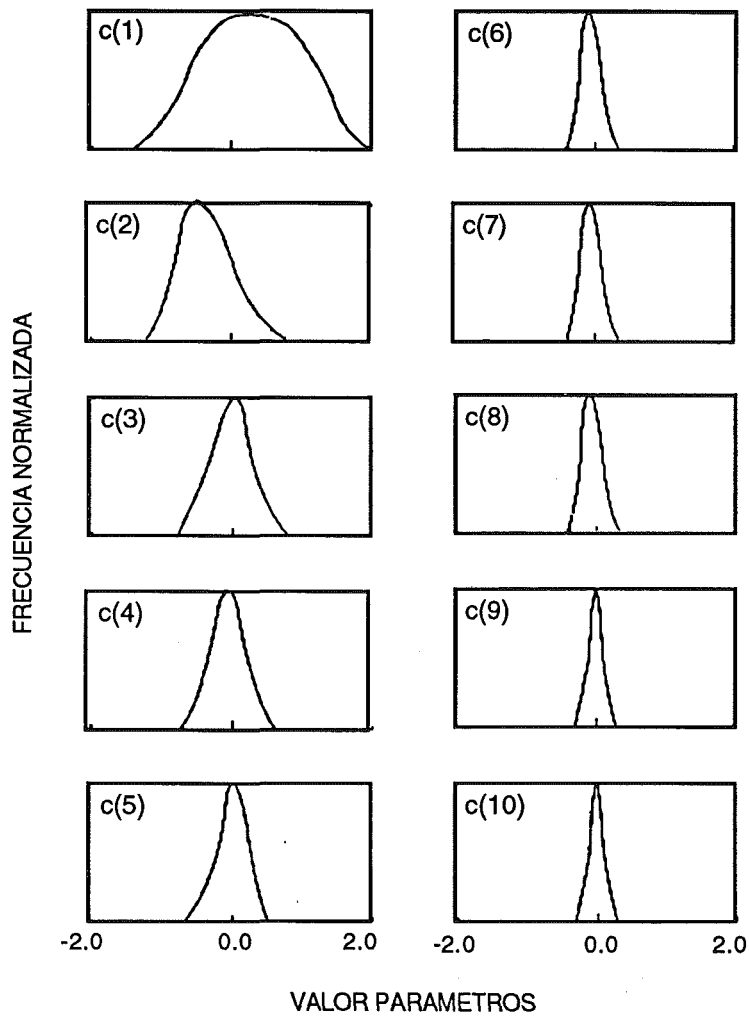

Fig. 4.2. Distribución estadística de los coeficientes cepstrum.

Dado un vector de características **c**, se puede expresar la probabilidad de que pertenezca a una clase i como P(i|**c**), la probabilidad de la clase i condicionada al vector **c**. Dadas dos clases cualesquiera i y k, se decide que el vector **c** pertenece a la case i, en lugar de a la clase k, si P(i|**c**) > P(k|**c**).

Aplicando la regla de Bayes, se tiene que

$$P(i|\mathbf{c}) = P(\mathbf{c}|i)\,\frac{P(i)}{P(\mathbf{c})}, \qquad (4.11)$$

donde $P(\mathbf{c}|i)$ es la probabilidad del vector $\mathbf{c}$ dada la clase i y P(i) y $P(\mathbf{c})$ son las probabilidades de la clase i y del vector $\mathbf{c}$, respectivamente. Por tanto, si se supone que las clases son equiprobables, P(i) = P(k), la regla de decisión consistirá en elegir la clase que maximiza $P(\mathbf{c}|i)$.

Denotando con $f_i(\mathbf{c})$ la función de densidad de probabilidad de la clase i evaluada en el vector $\mathbf{c}$, se tiene que $P(\mathbf{c}|i) > P(\mathbf{c}|k)$ si $f_i(\mathbf{c}) > f_k(\mathbf{c})$. La regla de decisión puede reformularse, pues, como elegir la clase que maximiza $f_i(\mathbf{c})$.

La función de densidad de probabilidad de la clase i, teniendo en cuenta el carácter gaussiano de las distribuciones, puede escribirse como

$$f_i(\mathbf{c}) = (2\pi)^{-m/2}\,|\mathbf{V}_i|^{-1/2}\,\exp\left[\left(-\frac{1}{2}\right)(\mathbf{c}-\bar{\mathbf{c}}_i)^T\mathbf{V}_i^{-1}(\mathbf{c}-\bar{\mathbf{c}}_i)\right], \qquad (4.12)$$

donde $\bar{\mathbf{c}}_i$ es la vector media y $\mathbf{V}_i$ la matriz de covarianza de la clase i y $|\mathbf{V}_i|$ denota el determinate de la matriz $\mathbf{V}_i$.

Para simplificar el cálculo puede maximizarse $\ln[f_i(\mathbf{c})]$, en lugar de $f_i(\mathbf{c})$, e ignorar el factor constante $(2\pi)^{-m/2}$. Por tanto, la función a maximizar será

$$\ln\left\{|\mathbf{V}_i|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{c}-\bar{\mathbf{c}}_i)^T\mathbf{V}_i^{-1}(\mathbf{c}-\bar{\mathbf{c}}_i)\right]\right\}. \qquad (4.13)$$

Eliminando el signo menos y el factor 1/2, el problema se reduce a minimizar

$$(\mathbf{c}-\bar{\mathbf{c}}_i)^T\mathbf{V}_i^{-1}(\mathbf{c}-\bar{\mathbf{c}}_i) + \ln|\mathbf{V}_i| \,. \qquad (4.14)$$

Este es el criterio básico de máxima verosimilitud. Sin embargo, raras veces es usado en su forma exacta. En primer lugar, los sistemas de reconocimiento se desarrollan sobre una base de datos limitada, que no permite una estimación fiable de las matrices $\mathbf{V}_i$. En segundo lugar, el uso del criterio de máxima verosimilitud en su forma exacta no es tan importante como una elección inteligente de los vectores de

características para obtener unos buenos resultados de reconocimiento. Por ello, en la práctica suele hacerse simplificaciones sobre este criterio.

Una posible simplificación es suponer que las matrices $\mathbf{V}_i$ de todas las clases son iguales a una matriz $\mathbf{V}$. De hecho, las matrices de covarianza para diferentes sonidos son suficientemente similares. Ello permite estimar esta matriz con un mayor número de datos. En este caso, el factor ln |V| es constante y puede ser ignorado. La expresión resultante es

$$(\mathbf{c}-\bar{\mathbf{c}}_i)^T \mathbf{V}^{-1}(\mathbf{c}-\bar{\mathbf{c}}_i) \ . \qquad (4.15)$$

Esta expresión da lugar a la siguiente medida de distancia entre dos vectores del espacio de características $\mathbf{c}_1$ y $\mathbf{c}_2$

$$d_M(\mathbf{c}_1,\mathbf{c}_2) = (\mathbf{c}_1-\mathbf{c}_2)^T \mathbf{V}^{-1}(\mathbf{c}_1-\mathbf{c}_2), \qquad (4.16)$$

que se conoce como distancia de Mahalanobis o distancia ponderada con covarianza. Esta distancia ya fue usada por Atal en 1976 [Ata76].

En el caso de un sistema de reconocimiento mediante comparación de patrones, los vectores $\mathbf{c}_1$ y $\mathbf{c}_2$ se corresponderán con los vectores de parámetros de test y referencia. En el caso de algunos tipos de modelos ocultos de Markov usados en las pruebas experimentales realizadas en este trabajo, discretos y de múltiple etiquetado, esta será la distancia aplicada para construir el diccionario del cuantificador vectorial y para cuantificar los vectores de características para obtener las etiquetas. En la primera situación $\mathbf{c}_1$ y $\mathbf{c}_2$ serán dos vectores cualesquiera de la base de datos utilizada en la construcción del cuantificador; en la segunda situación, uno de ellos será el vector a cuantificar y el otro será cada una de las palabras-código del diccionario del cuantificador.

En el proceso de construcción del diccionario de un cuantificador vectorial el número de veces que se calcula la distancia entre pares de vectores de la base es extraordinariamente alto. Si la distancia es muy simple, como la euclídea, el tiempo de cálculo es razonable. Sin embargo, en el caso de la distancia de Mahalanobis el proceso es extremadamente lento.

Por ello, en las pruebas experimentales realizadas en este trabajo, la distancia de Mahalanobis se ha implementado de un modo eficiente para poder construir el diccionario del cuantificador vectorial de los modelos discretos de Markov en un tiempo razonable. Esta implementación ha consistido en una transformación lineal de los vectores de parámetros de tal forma que la distancia de Mahalanobis sobre los vectores originales equivale a la distancia euclídea sobre los vectores transformados.

Debido a que la matriz de covarianza **V** es simétrica y definida positiva, se puede realizar la siguiente descomposición

$$\mathbf{V} = \mathbf{U}\,\mathbf{D}\,\mathbf{U}^T, \qquad (4.17)$$

donde **U** es una matriz unitaria formada por los vectores propios de **V**, que son ortonormales entre sí, y **D** es una matriz diagonal formada por los valores propios asociados, que son positivos.

Aprovechando esta descomposición, si se realiza la siguiente transformación lineal sobre los vectores cepstrales

$$\mathbf{c}_i' = \mathbf{D}^{-1/2}\,\mathbf{U}\,\mathbf{c}_i \qquad i = 1, 2, \qquad (4.18)$$

donde $\mathbf{D}^{-1/2}$ es una matriz diagonal formada por los inversos de las raíces cuadradas positivas de los elementos de **D**, es fácil demostrar que la distancia de Mahalanobis entre los vectores originales $\mathbf{c}_i$ es equivalente a la distancia euclídea d_E entre los vectores transformados $\mathbf{c}_i'$, es decir,

$$d_M(\mathbf{c}_1,\mathbf{c}_2) = d_E(\mathbf{c}_1',\mathbf{c}_2') = (\mathbf{c}_1'-\mathbf{c}_2')^T(\mathbf{c}_1'-\mathbf{c}_2') \qquad (4.19)$$

(estrictamente la distancia euclídea es la raíz cuadrada de esta magnitud, pero en lo sucesivo no se tendrá en cuenta este hecho, ya que la raíz cuadrada aumenta el coste de cálculo sin variar las prestaciones del sistema).

Una vez aplicada la expresión (4.18) a todos los vectores que se desean utilizar para construir el diccionario del cuantificador vectorial, se puede aplicar la distancia euclídea en dicha construcción. Las palabras-código resultantes corresponden al espacio de características transformado. Por tanto, en el proceso de cuantificación

basta con transformar los vectores a cuantificar y utilizar la distancia euclídea para encontrar la palabra-código correspondiente.

4.3. DISTANCIAS EUCLIDEAS CEPSTRALES PONDERADAS

La definición del cepstrum dada en (4.1) puede escribirse de la forma

$$\ln S(\omega) = \sum_{n=-\infty}^{\infty} c(n)e^{-j\omega n} \qquad (4.20)$$

y, teniendo en cuenta la simetría del espectro, también puede escribirse como

$$\ln S(\omega) = 2 \sum_{n=1}^{\infty} c(n) \cos(\omega n) + c(0). \qquad (4.21)$$

Esta serie puede aproximarse mediante una suma finita, llamada transformada coseno discreta (DCT). Puede demostrarse [Zel77] que la base de funciones intrínsecas a la DCT tienen una estrecha semejanza con los vectores propios de la transformación óptima de Karhunen-Loève (KLT). Por tanto, la representación DCT del espectro aproxima bastante bien la descomposición ortogonal óptima KLT. De acuerdo con esta conclusión, se observa experimentalmente que la matriz de covarianza de los coeficientes cepstrales es predominantemente diagonal, es decir los coeficientes cepstrales están bastantes incorrelados entre sí.

Debido a que los elementos de fuera de la diagonal de la matriz de covarianza de los coeficientes cepstrales son relativamente pequeños con respecto a los de la diagonal, existen problemas de precisión en la estimación de los elementos de fuera de la diagonal de la matriz. A este problema, se añade la sensibilidad introducida por la inversión de la matriz necesaria para calcular la distancia de Mahalanobis y la complejidad computacional inherente a la distancia.

Teniendo en cuenta estos factores, una simplificación muy usada en la práctica es considerar diagonal la matriz de covarianza. Ello es consistente con el hecho de que los coeficientes cepstrales están bastante incorrelados entres sí y soluciona los problemas de implementación mencionados.

De este modo, la distancia de Mahalanobis se convierte en la siguiente distancia euclídea cepstral ponderada [Fur81]

$$d_{WE}(c_1,c_2)= \sum_{n=1}^{L} \left(\frac{1}{\sigma_n}\left(c_1(n)-c_2(n)\right)\right)^2 , \qquad (4.22)$$

donde σ_n es la desviación típica del coeficiente cepstral c(n). Por ello, en esta memoria esta distancia se denotará con el nombre de distancia euclídea cepstral ponderada con la inversa de la desviación típica. Tanto en esta distancia como en el resto de distancias euclídeas cepstrales ponderadas descritas en este apartado, siempre se ha de cumplir que la longitud del vector cepstral L sea mayor o igual que el orden del modelo de predicción lineal p para que la distancia sea definida positiva.

Otra forma más directa de derivar la distancia (4.22) consiste simplemente en considerar el hecho de que la desviación típica de los coeficientes cepstrales decae monótonamente con el índice (ver figura 4.2). Por tanto, si los coeficientes cepstrales de orden alto llevan información útil para el reconocimiento de voz no estarían utilizados con efectividad en el caso de usar una simple medida de distancia euclídea.

En [Toh87] se muestra que esta distancia tiene un buen comportamiento y que los mejores resultados de reconocimiento se obtienen cuando se utiliza una longitud de vector cepstral L igual al orden del modelo de predicción lineal p. También se muestra que este buen comportamiento es debido principalmente al desénfasis de los coeficientes cepstrales de orden bajo, debido a su gran sensibilidad a las características del locutor y el canal de transmisión, más que a la ponderación de los coeficientes de orden alto. Si se utiliza un valor de L superior a p los resultados de reconocimiento bajan considerablemente. Por ello, se recomienda otro tipo de ponderación de los coeficientes de índice mayor que p si se quiere hacer uso de ellos en reconocimiento.

Si se considera que todas las desviaciones típicas de los coeficientes cepstrales son iguales, la distancia (4.22) queda reducida a la distancia euclídea cepstral

$$d_E(c_1,c_2)= \sum_{n=1}^{L} \left(c_1(n)-c_2(n)\right)^2 , \qquad (4.23)$$

Esta distancia puede considerarse como una implementación eficiente de la distancia euclídea sobre el logaritmo del espectro

$$d_E\,(\ln S_1,\ln S_2) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln S_1(\omega)-\ln S_2(\omega)\right)^2 d\omega, \qquad (4.24)$$

una de las primeras distancias propuestas en procesado de voz y que tiene una justificación desde el punto de vista perceptivo teniendo en cuenta que el oído humano realiza una comprensión logarítmica en intensidad para cada frecuencia. Esta distancia puede considerarse, además, como un caso particular de las distancias L_p, para $p = 2$ [Gra76]; sin embargo, se comprueba que el valor de p es poco importante en cuanto al poder discriminativo de la distancia y que el coste de cálculo para un valor de p distinto de 2 es elevado.

Para observar la relación entre las distancias (4.23) y (4.24) basta utilizar la relación de Parseval y ciertas propiedades de los coeficientes cepstrales. Si se aplica sobre la expresión (4.24) la relación de Parseval y se tiene en cuenta el carácter par de la secuencia cepstrum, se obtiene

$$d_E\,(S_1,S_2) = 2\sum_{n=1}^{\infty}\left(c_1(n)-c_2(n)\right)^2 + \left(c_1(0)-c_2(0)\right)^2 \qquad (4.25)$$

Considerando que el término del cepstrum en el origen es el logaritmo de la ganancia del modelo y no interesante de cara al reconocimiento y que son suficientes un número no excesivo de términos de la secuencia cepstral para aproximar el modelo espectral, la serie (4.25) queda reducida a la expresión (4.23) si se ignora el factor 2.

Por otro lado, una clase de distancias que ha sido considerada en reconocimiento automático del habla es la basada en la derivada del logaritmo del espectro debido a su alta correlación con distinciones fonéticas subjetivas y sus buenas prestaciones. Klatt [Kla82] fue el primero en realizar estudios psicoacústicos en este sentido y proponer una distancia de ese tipo, llamada en inglés *Weighted Slope Metric*, aplicada sobre el espectro estimado mediante un banco de filtros correspondientes a las bandas críticas del oído.

Hanson y Wakita [Han87] aplicaron este concepto al espectro correspondiente al modelo de predicción lineal y comprobaron que la distancia euclídea entre las derivadas de los logaritmos de los espectros

$$d_E\left(\frac{d}{d\omega}\ln S_1, \frac{d}{d\omega}\ln S_2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{d}{d\omega}\ln S_1(\omega) - \frac{d}{d\omega}\ln S_2(\omega)\right)^2 d\omega \qquad (4.26)$$

es muy sensible a variaciones en los picos de los formantes. Debido a que los picos de los formantes son las zonas de mayor energía del espectro y las menos afectadas por el ruido aditivo, esta medida distancia es una buena candidata para el reconocimiento en entornos ruidosos. Por otro lado, teniendo en cuenta la consistencia de esta distancia con los estudios perceptivos realizados, también es de esperar su buen comportamiento en reconocimiento robusto.

El principal problema de la distancia (4.26) es su elevado coste de cálculo. Sin embargo, repitiendo los mismos razonamientos que han relacionado las expresiones (4.23) y (4.24), se puede proponer como una implementación eficiente de esta distancia la siguiente distancia euclídea cepstral ponderada

$$d_{WE}(\mathbf{c}_1,\mathbf{c}_2) = \sum_{n=1}^{L}\left(n\left(c_1(n)-c_2(n)\right)\right)^2 . \qquad (4.27)$$

Los coeficientes cepstrales ponderados con su índice nc(n) utilizados en la fórmula anterior han sido referenciados como *root-power sums* (RPS) [Sch81] o coeficientes cepstrales ponderados en la cuefrencia [Pal82].

En la práctica, esta distancia no difiere mucho de la distancia euclídea cepstral ponderada con la inversa de cada coeficiente, ya que puede demostrarse, bajo ciertas condiciones, que los coeficientes cepstrales de índice mayor que cero tienen desviaciones típicas inversamente proporcionales al índice aproximadadmente, es decir,

$$\sigma_n \propto \frac{1}{n} . \qquad (4.28)$$

Debido a ello, al igual que en el caso de la distancia euclídea cepstral ponderada con la inversa de la desviación típica, en [Han87] se muestra que esta distancia tiene un buen comportamiento y que los mejores resultados de reconocimiento se obtienen cuando se utiliza una longitud de vector cepstral L igual al orden del modelo de

predicción lineal p. Si se utiliza un valor de L superior a p los resultados de reconocimiento bajan considerablemente. También se muestra que este buen comportamiento es debido principalmente al desénfasis de los coeficientes cepstrales de orden bajo.

En el caso de reconocimiento de habla en presencia de ruido de espectro plano, el desénfasis de los coeficientes cepstrales de orden bajo puede resultar muy beneficioso, pues estos son los coeficientes más contaminados por el ruido. Otra implementación de la distancia sobre la pendiente espectral fue propuesta por Stanton para el reconocimiento del habla en ambiente ruidoso [Sta89].

La distancia euclídea cepstral ponderada con la inversa de la desviación típica (4.22), la distancia euclídea cepstral sin ponderación (4.23) y esta última distancia (4.27), pueden considerarse como casos particulares de la distancia euclídea cepstral ponderada genérica [Pal82]

$$d_{WE}(\mathbf{c}_1,\mathbf{c}_2)= \sum_{n=1}^{L} \left(w(n) \left(c_1(n)-c_2(n)\right)\right)^2 , \qquad (4.29)$$

donde w(n) es una función de ponderación. En la práctica, en primer lugar se ponderan los coeficientes cepstrales con la función w(n) pertinente y luego se aplica la distancia euclídea cepstral sin ponderación (4.23) sobre los coeficientes cepstrales ponderados. Por ello, a la función w(n) se le suele llamar ventana de ponderación cepstral o liftado.

En el caso de la distancia euclídea cepstral sin ponderación, esta ventana es simplemente

$$w(n) = 1 \qquad n = 1,\ldots L. \qquad (4.30)$$

Como en las otras dos distancias, el valor óptimo de L es el orden de predicción lineal p, la ventana correspondiente a la distancia euclídea cepstral ponderada con la inversa de la desviación típicas es

$$w(n) = \frac{1}{\sigma_n} \qquad n = 1,\ldots L = p \qquad (4.31)$$

y la correspondiente a la distancia (4.27)

$$w(n) = n \qquad n = 1,\ldots L = p. \qquad (4.32)$$

Por ello a esta última distancia se le denominará en esta memoria distancia euclídea cepstral ponderada con la ventana rampa.

Tanto en el caso de la distancia euclídea cepstral ponderada con la inversa de la desviación típica como en el de la ponderada con la ventana rampa, ya se ha comentado que su buen comportamiento es debido al desénfasis de los coeficientes cepstrales de orden bajo más que al énfasis de los de orden alto. Concretamente, la longitud óptima de la ventana de ponderación L coincide con el orden de predicción lineal p y los resultados de reconocimiento empeoran al aumentar la longitud de L por encima del valor de p debido a la ponderación excesiva de los coeficientes cepstrales de orden alto. Si se quiere utilizar estos coeficientes cepstrales es necesario emplear otro tipo de ponderación.

Juang [Jua87a] realizó un estudio sobre las fuentes de variabilidad de los coeficientes cepstrales LPC y concluyó que tanto los coeficientes cepstrales de orden bajo como los de orden alto presentan variabilidades no deseadas para el reconocimiento. De un lado, constató que los coeficientes de orden bajo presentan una variabilidad provocada por las condiciones del canal de transmisión y las características propias del locutor tales como la forma de onda glotal. En cuanto a los coeficientes de orden alto, comprobó mediante estudios de simulación que presentan una variabilidad artificialmente alta debida al proceso de análisis.

De este conjunto de consideraciones, se deduce que la función de ponderación de los coeficientes cepstrales debe desenfatizar tanto los coeficientes de orden bajo como los de orden alto. Tras diversas pruebas experimentales utilizando funciones de ponderación de este tipo para orden de predicción p igual a 8, se Juang propuso la siguiente ventana cepstral

$$w(n) = 1 + \frac{L}{2}\,\text{sen}\left(\frac{\pi n}{L}\right) \qquad n = 1,\ldots \frac{3}{2}p \qquad (4.33)$$

Esta ventana en forma de seno realzado (*bandpass liftering*, en la literatura inglesa) ha sido ampliamente utilizada en reconocimiento automático del habla.

En la figura 4.3 se compara la forma de las ventanas inversa de la desviación típica, rampa y seno realzado, normalizadas respecto al valor de la ventana en n = 1. La ventana inversa de la desviación típica ha sido estimada utilizando la base de datos empleada en las pruebas experimentales realizadas en este trabajo con ruido blanco. En este caso, esta ventana queda globalmente por debajo de la ventana rampa.

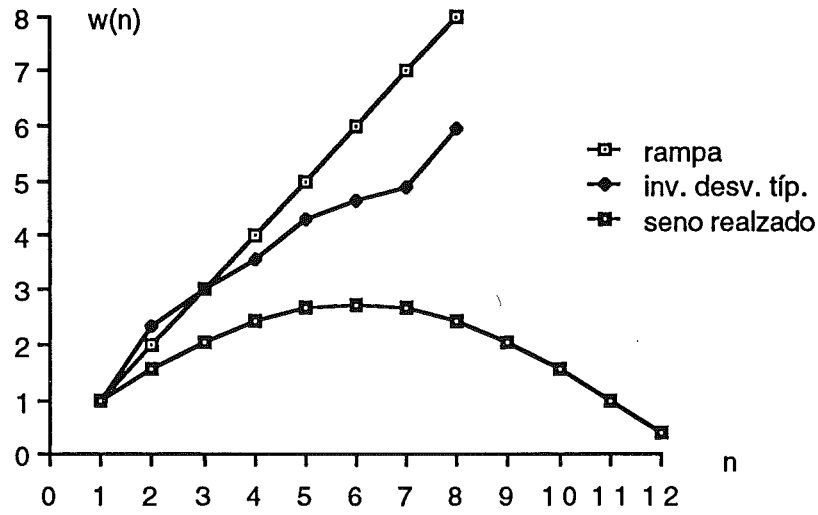

Fig. 4.3. Comparación de las ventanas inversa de la desviación típica, rampa y seno realzado para orden de predicción p igual a 8.

La aplicación de cualquiera de estas ventanas sobre la secuencia de cepstrum infinita da lugar a una versión suavizada del logaritmo del espectro que depende de la forma y longitud de la ventana de ponderación. Uno de los objetivos de este trabajo será, por tanto, obtener el grado óptimo de suavizado en condiciones ruidosas.

Hay que hacer notar que este nuevo espectro suavizado ya no se corresponde con un modelo autorregresivo sino con un modelo de cepstrum finito, que no preserva los valores iniciales de la autocorrelación. En la figura 4.4. se observa la secuencia de espectros LPC de orden de predicción lineal 8 correspondientes a una pronunciación del dígito "dos" en catalan (a) y la misma secuencia suavizada mediante una ventana cepstral rectangular de longitud 8 (b), una ventana seno realzado de longitud 12 (c) y una ventana rampa de longitud 8.

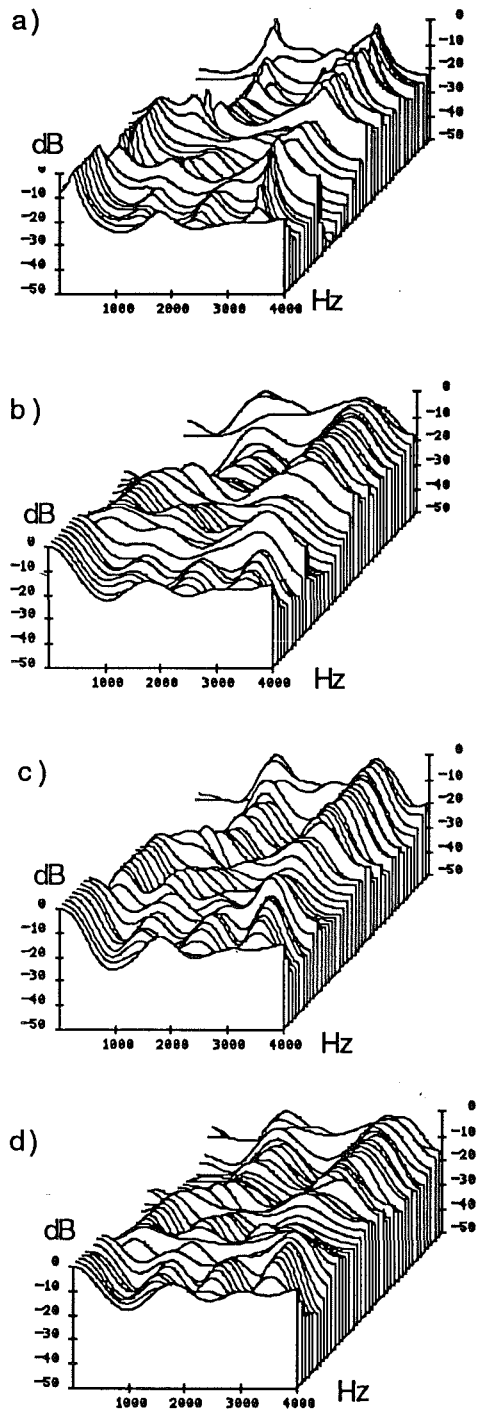

Fig. 4.4. Influencia de distintas ventanas cepstrales sobre una secuencia de espectros LPC de orden 8: a) sin ventana b) rectangular c) seno realzado d) rampa

En la figura puede observarse la presencia de picos abruptos y la variabilidad de las medidas espectrales que caracterizan a los espectros LPC sin suavizar cepstralmente. También se muestra que en las versiones suavizadas se reduce esta variabilidad, preservando las características esenciales de la estructura de formantes.

La medida de distancia de retardo de grupo suavizado (SGD, *Smoothed Group Delay*) fue propuesta para el reconocimiento en presencia de ruido y distorsiones [Ita87] y no es más que un caso particular de distancia euclídea cepstral ponderada en la que la función de ponderación se define como

$$w(n) = n^s \exp\left(\frac{n^2}{2\tau^2}\right) \qquad s \geq 0 \qquad (4.34)$$

Experimentos con habla ruidosa y distorsionada han indicado que la ponderación más efectiva se obtiene con valores de s entre 1 y 2 y τ alrededor de 5. Es interesante destacar que con estos valores de los parámetros la función de ponderación SGD es muy similar a la ponderación seno realzado propuesta en [Jua87a].

Por último, hay que hacer notar que, a diferencia de otras medidas de distorsión espectral, las distancias euclídeas cepstrales ponderadas tienen una interpretación geométrica sencilla y un coste de cálculo muy pequeño, lo cual las hace muy apropiadas como medidas de distorsión en el proceso de cuantificación vectorial inherente a ciertos tipos de modelos de Markov que se utilizarán en las pruebas experimentales de esta memoria.

4.4. DISTANCIAS DE PROYECCION CEPSTRAL

En el caso de reconocimiento de habla ruidosa utilizando referencias libres de ruido, puede argumentarse que no existe justificación para mantener la simetría que presentan las distancias euclídeas cepstrales ponderadas revisadas en el apartado anterior. Así, por ejemplo, desde el punto de vista de la percepción del oído humano, el enmascaramiento del habla por el ruido es un fenómeno distinto al enmascaramiento del ruido por el habla, por lo que una medida de distorsión asimétrica podría ser más realista.

A partir de estudios analíticos, confirmados experimentalmente, Mansour y Juang [Man89b] mostraron que el ruido gaussiano aditivo reduce la norma del vector cepstral, construido con los primeros coeficientes cepstrales c(n) de n = 1 a L, y que esta reducción es función de la relación señal-ruido; a menor relación señal-ruido, mayor reducción de la norma. Además, a través de histogramas, los mismos autores comprobaron que los vectores cepstrales de mayor norma son más robustos al ruido que los vector de norma menor y que el ángulo entre el vector cepstral limpio y su versión ruidosa es especialmente robusto al ruido.

Teniendo en cuenta estos efectos y la no necesidad de mantener la simetría de la distancia euclídea cepstral en reconocimiento de habla ruidosa, Mansour y Juang propusieron una familia de distancias basadas en la proyección entre vectores cepstrales ponderados para su utilización en un sistema de reconocimiento automático del habla mediante comparación de patrones cuando el sistema es entrenado en condiciones libres de ruido pero las condiciones de test son ruidosas.

Por otro lado, también constataron el hecho bien conocido de que los coeficientes cepstrales de orden bajo son más sensibles al ruido que los de orden alto. Para compensar este efecto utilizaron la siguiente ventana de ponderación cepstral sobre el cepstrum LPC de orden de predicción 8

$$w(n) = 0 \qquad n = 1 \qquad (4.35)$$

$$w(n) = 1 + 6.5 \operatorname{sen}\left(\frac{\pi n}{L}\right) \qquad n = 2,\ldots 12, \qquad (4.36)$$

que suele denominarse ventana seno desplazado, por comparación con la ventana seno realzado (4.33).

Para compensar el efecto perjudicial de la disminución de la norma del vector de test ruidoso cuando se utiliza la distancia euclídea cepstral, un método simple consiste en realizar una ecualización de primer orden usando la siguiente medida de distancia

$$d_{P1}(\mathbf{c}_t,\mathbf{c}_r) = (\mathbf{c}_t - \lambda\, \mathbf{c}_r)^T(\mathbf{c}_t - \lambda\, \mathbf{c}_r), \qquad (4.37)$$

donde $\mathbf{c}_t$ es el vector cepstral ponderado de test ruidoso, $\mathbf{c}_r$ es el vector de referencia limpio y λ una factor de ponderación que depende de la relación señal-ruido global.

Para cada relación señal-ruido existe un valor óptimo de λ que proporciona los mejores resultados de reconocimiento.

Una posible manera de ecualizar la norma de los vectores sin un conocimiento previo de la relación señal-ruido es elegir el valor de λ que minimiza la distancia d_{P1} para cada trama. A partir del principio de ortogonalidad, este valor óptimo satisface la siguiente ecuación

$$(\mathbf{c}_t - \lambda_{opt}\,\mathbf{c}_r)^T \mathbf{c}_r = 0, \tag{4.38}$$

cuya solución es

$$\lambda_{opt} = \frac{\mathbf{c}_t^T \mathbf{c}_r}{\mathbf{c}_r^T \mathbf{c}_r} \tag{4.39}$$

Sustituyendo (4.39) en (4.37), resulta la siguiente distancia optimizada

$$d_{P2}(\mathbf{c}_t,\mathbf{c}_r) = |\mathbf{c}_t|^2 (1 - \cos^2 \beta), \tag{4.40}$$

donde cos β se define como

$$\cos\beta = \frac{\mathbf{c}_t^T \mathbf{c}_r}{|\mathbf{c}_t|\,|\mathbf{c}_r|}, \tag{4.41}$$

es decir, el ángulo o coseno direccional entre los dos vectores comparados.

Estas dos distancias han sido derivadas teniendo en cuenta únicamente el efecto de reducción de la norma. Por otro lado, considerando la especial robustez del ángulo entre el vector cepstral limpio y su versión ruidosa, la distancia euclídea entre las versiones normalizadas de los dos vectores cepstrales resulta ser también una distancia robusta

$$d_{P3}(\mathbf{c}_t,\mathbf{c}_r) = \left(\frac{\mathbf{c}_t}{|\mathbf{c}_t|} - \frac{\mathbf{c}_r}{|\mathbf{c}_r|}\right)^T \left(\frac{\mathbf{c}_t}{|\mathbf{c}_t|} - \frac{\mathbf{c}_r}{|\mathbf{c}_r|}\right) = 2\,(1 - \cos\beta). \tag{4.42}$$

Si se considera, además, que los vectores cepstrales con normas mayores son más robustos al ruido blanco aditivo que los vectores de normas más pequeñas, la misma norma del vector de test puede usarse como ponderación en la medida de distancia para enfatizar aquellas tramas de la señal de test que son más fiables en la distancia acumulada calculada por el algoritmo de programación dinámica. Aplicando esta idea sobre la distancia d_{P3}, resulta la siguiente medida de distancia

$$d_{P4}(\mathbf{c}_t,\mathbf{c}_r) = |\mathbf{c}_t|^{\alpha}(1-\cos\beta) \tag{4.43}$$

en función del parámetro α. Para $\alpha = 0$, las distancias d_{P3} y d_{P4} coinciden. Para $\alpha = 1$, se obtiene

$$d_{P5}(\mathbf{c}_t,\mathbf{c}_r) = |\mathbf{c}_t|\,(1-\cos\beta), \tag{4.44}$$

que sustituyendo el valor de cos β, puede escribirse como

$$d_{P5}(\mathbf{c}_t,\mathbf{c}_r) = |\mathbf{c}_t| - \frac{\mathbf{c}_t^T\mathbf{c}_r}{|\mathbf{c}_r|}, \tag{4.45}$$

Como en el algoritmo de programación dinámica puede sumarse una constante a la medida de distancia para cada trama de test sin alterar los resultados, d_{P5} puede escribirse de la forma

$$d_{P5}(\mathbf{c}_t,\mathbf{c}_r) = -\,\mathbf{c}_t^T\hat{\mathbf{c}}_r, \tag{4.46}$$

donde $\hat{\mathbf{c}}_r$ es el vector de referencia limpio normalizado.

La distancia d_{P5} tiene la forma de una operación de proyección entre el vector de test y el vector de referencia normalizado y tiene el mismo coste de cálculo que la distancia euclídea. Esta distancia y las distancias descritas anteriormente que utilizan operaciones de proyección forman una familia de distancias conocidas como distancias de proyección cepstrales.

En pruebas de reconocimiento en ambiente ruidoso, estas distancias han mostrado un comportamiento superior a las distancias euclídeas cepstrales ponderadas,

tanto en sistemas que utilizan comparación de patrones [Man89b] como en sistemas basados en modelos ocultos de Markov continuos [Car91].

En presencia de efecto Lombard también se produce una reducción de la norma del vector cepstral, por lo cual estas distancias proporcionan buenos resultados [Jun89].

4.5. INCORPORACION DE INFORMACION DINAMICA Y ENERGIA

Las características espectrales dinámicas de la señal de voz juegan un papel importante en la percepción humana de la voz [Rus82], siendo las zonas de la señal donde la variación espectral es máxima las que aportan la mayor cantidad de información fonética [Fur84]. Por otro lado, la informaciones del espectro instantáneo y su derivada son complementarias [Rab88], resultando esta última más robusta a la variabilidad interlocutor y del entorno [Fur86]. Por todas estas razones, en la mayoría de los sistemas actuales de reconocimiento automático del habla se incorpora información dinámica del espectro para mejorar sus prestaciones.

Si se denota como c(n,t) la secuencia cepstral, referida hasta ahora como c(n), estimada a partir de la trama de señal correspondiente al instante t, la variación con el tiempo del logaritmo del espectro viene dada por

$$\frac{\partial(\ln S(\omega,t))}{\partial t} = \sum_{n=-\infty}^{\infty} \frac{\partial c(n,t)}{\partial t} e^{-jn\omega}, \qquad (4.47)$$

siendo S(ω,t) el espectro estimado en el instante t.

La estimación de la derivada de c(n,t) con respecto al tiempo t puede realizarse simplemente restando dos valores separados por una espacio de tiempo adecuado [Shi86b] [Gup87] [Lee88a] [App90], es decir

$$d(n,t) = c(n, t+\delta) - c(n, t-\delta). \qquad (4.48)$$

Sin embargo, la diferencia finita de primer orden es intrínsecamente ruidosa.

Otra posibilidad más difundida [Fur86] [Jun87] [Rab88] [Han90], que no presenta el inconveniente anterior, consiste en aplicar regresión lineal sobre la evolución temporal de cada coeficiente cepstral en un intervalo de duración adecuada. En este caso, la estimación de la derivada viene dada por el coeficiente de regresión lineal $\Delta c(n,t)$

$$\Delta c(n,t) = \frac{\sum_{k=-K}^{K} k\, c(n,t+k)}{\sum_{k=-K}^{K} k^2}, \qquad (4.49)$$

donde donde K define el intervalo de estimación, desde t-K a t+K.

Estos coeficientes $\Delta c(n,t)$ son conocidos también con el nombre de delta-cepstra y son una estimación de la rapidez de variación de la función temporal de cada parámetro en cada segmento. En concreto, se corresponden con la pendiente del polinomio de orden uno que mejor se aproxima a esta función temporal minimizando el error cuadrático en dicho intervalo.

La longitud del intervalo de regresión debe elegirse suficientemente grande para obtener estimaciones adecuadas de las características dinámicas del espectro y suficientemente breve para que no se introduzca un suavizado excesivo en los valores estimados, de forma que los parámetros dinámicos modelen adecuadamente las zonas transicionales de la señal producidos por los efectos coarticulatorios entre fonemas. Se suelen utilizar intervalos de una duración comprendida entre 50 y 110 ms.

Por otro lado, se puede interpretar la expresión (4.49) como un filtrado paso-alto de los coeficientes cepstrales. Se han realizado varios estudios en este sentido, utilizando filtrado paso-banda [Her91] o paso-alto [Hir91] de los componentes espectrales para combatir las fuentes de variabilidad de la señal de voz. Concretamente, Hirsch propuso en [Hir91] el siguiente filtro paso-alto IIR de primer orden sobre los componentes espectrales

$$y(n) = x(n) - x(n-1) + 0.7\, y(n-1), \qquad (4.50)$$

que tiene una frecuencia de corte aproximadamente en 4.5 Hz.

Los valores $\Delta c(n,t)$ constituyen un nuevo vector de parámetros acústicos. Por razonamientos análogos a los expuestos en el apartado 4.3 para el caso del vector cepstral, para calcular la similitud entre estos vectores suele utilizarse una distancia euclídea ponderada

$$d_{WE}(\Delta \mathbf{c}_1,\Delta \mathbf{c}_2)= \sum_{n=1}^{L} \left(w(n) \left(\Delta c_1(n,t)-\Delta c_2(n,t)\right)\right)^2 , \qquad (4.51)$$

donde $\Delta \mathbf{c}_1$ y $\Delta \mathbf{c}_2$ son los vectores delta-cepstrales a comparar.

Por otro lado, también se ha comprobado que la utilización de derivadas de orden superior de las funciones temporales de la coeficientes cepstrales mejora los resultados obtenidos en reconocimiento automático del habla tanto limpia como ruidosa. Ello es debido a que así se obtiene una representación más completa bidimensional (tiempo y frecuencia) de la señal de voz.

Para la estimación de estas derivadas se han propuesto fundamentalmente dos alternativas: el uso de coeficientes de regresión de orden superior [Han90] y el cálculo iterativo del coeficiente de regresión de primer orden [Ney90]. A los parámetros de segundo orden obtenidos utilizando esta segunda alternativa se les suele denominar delta-delta-cepstra.

Otro parámetro que se ha mostrado útil en reconocimiento automático del habla es el logaritmo de la energía de cada trama de señal de voz [Bro82] [Rab84]. La transformación logarítmica aplicada a la energía local la aproxima a la escala perceptivo de sonoridad. En esta memoria cuando se hable de energía se considerará implícita dicha transformación logarítmica.

Sin embargo, la energía, por sí misma, no es una fuente fiable de información ya que presenta grandes variaciones entre locutores e incluso para un mismo locutor debidas a que es función del volumen de la locución. Por ello es necesario realizar algún tipo de normalización sobre la misma. La aproximación más simple consiste en restar a la energía de cada trama el valor máximo de esta en la palabra. Sin embargo, cuando se ha de trabajar en tiempo real, el esperar a detectar el final de la palabra para empezar a formar los vectores de características y procesarlos es inviable.

También se ha mostrado útil en reconocimiento del habla [Fur86] el coeficiente de regresión lineal de la energía, o delta-energía, $\Delta E(t)$

$$\Delta E(t) = \frac{\sum_{k=-K}^{K} k\, E(t+k)}{\sum_{k=-K}^{K} k^2}, \qquad (4.52)$$

que representa la evolución dinámica de la energía de la señal. Este parámetro tiene la ventaja con respecto a la energía logarítmica de que no necesita ser normalizado, debido a que está expresado en términos de diferencias de energías logarítmicas.

Por último, también se han propuesto parámetros dinámicos de la energía de orden superior, como el delta-delta-energía [Wil91], para mejorar las prestaciones de los sistemas de reconocimiento.

En las pruebas experimentales de este trabajo se analizará el comportamiento de todos estos parámetros en reconocimiento de habla ruidosa, variando el número de parámetros y el algoritmo e intervalo de estimación de los mismos.

También se considerarán las dos formas de incorporar estas informaciones a un sistema de reconocimiento basado en modelos ocultos de Markov con cuantificación vectorial:

- Distancia compuesta, que consiste en construir un supervector concatenando con una ponderación adecuada los vectores y/o las componentes escalares que se desean utilizar y obtener un único símbolo correspondiente a este supervector usando distancia euclídea ponderada en el proceso de cuantificación vectorial.

- Diccionarios múltiples, en la que se cuantifican por separado cada una de las informaciones y se considera independencia estadística de las mismas en el entorno de los modelos ocultos de Markov.