Estudio comparativo de varias representaciones paramétricas para el reconocimiento automático del habla

DEE-I/2/94

Luis Javier Rodríguez Fuentes
Departamento de Electricidad y Electrónica.
Universidad del País Vasco.
Apartado 644. 48080 Bilbao. SPAIN.
e-mail: luisja@we.lc.ehu.es

Noviembre 1994

GRAH-MBAT

Grupo de Reconocimiento Automático del Habla
Mintzo-Berehizkuntza Automatikoaren Taldea

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

Departamento de Electricidad y Electrónica
Elektrika eta Elektronika Saila

Agradecimientos

Quiero hacer constar mi agradecimiento a Inés Torres por sus sugerencias en el desarrollo de los experimentos y por la ayuda prestada en la redacción de esta memoria, y al Grupo de Reconocimiento de Formas e Inteligencia Artificial de la Universidad Politécnica de Valencia, especialmente a Jose Miguel Benedí y Enrique Vidal, por su activa participación en el diseño del trabajo y por la base de datos DIG2, sobre la cual se han llevado a cabo las pruebas experimentales.

Resumen

El objetivo de este trabajo es evaluar el rendimiento de un amplio conjunto de parametrizaciones mediante una tarea de reconocimiento sencilla. Se han comparado dos grupos de parámetros. Por un lado, parámetros derivados de análisis de predicción lineal: coeficientes de reflexión (RC), log-area ratios (LAR) y coeficientes cepstrales (LPCEP). En segundo lugar, se han considerado parámetros derivados de análisis en el dominio de la frecuencia: coeficientes cepstrales obtenidos a partir de una transformada rápida de Fourier (FFTCEP), coeficientes de un banco de filtros con escala Bark (BFB) y coeficientes cepstrales obtenidos a partir de dicho banco de filtros (BFBCEP).

Habitualmente se utilizan parámetros derivados de un banco de filtros con escala Bark cuando la frecuencia de muestreo es 16 kHz, mientras que los parámetros derivados de un análisis de predicción lineal se utilizan con frecuencias de muestreo inferiores (8, 10 kHz). Sin embargo, no nos constan evidencias experimentales a 16 kHz que justifiquen esta elección. Trataremos de resolver esta cuestión comparando directamente el rendimiento de ambas representaciones sobre una tarea de reconocimiento de palabras aisladas.

Tampoco se tiene constancia de cómo influye el preénfasis de la señal de voz en la extracción de características acústicas para reconocimiento, por lo que se va a evaluar el rendimiento de los parámetros con y sin preénfasis. Finalmente se introducirá la transformación bilineal de las secuencias de parámetros FFTCEP y LPCEP, con objeto de reproducir la escala Bark de bandas críticas.

El primer capítulo del informe presenta una visión general del problema del reconocimiento del habla y destaca la importancia de extraer características acústicas de calidad. En el capítulo 2 se describen aspectos relativos a la definición y el cálculo de los parámetros. Los capítulos 3 y 4 presentan la metodología de evaluación y los datos de la implementación. En el capítulo 5 se resumen y comentan brevemente los resultados experimentales. El informe finaliza con un capítulo de conclusiones y con una lista de referencias bibliográficas.

Abstract

The aim of this work is to evaluate a large set of parametric representations over an easy recognition task. Two groups of parametric representations are compared. The first one, Linear Prediction analysis (LP), includes reflection coefficients (RC), log-area ratios (LAR) and cepstral coefficients (LPCEP). The second group includes the Fast Fourier Transform derived cepstral coefficients (FFTCEP), the Bark-scaled Filter Bank coefficients (BFB) and the corresponding cepstral coefficients (BFBCEP).

The BFB derived parameters have been typically used for speech recognition with sampling rates of 16 kHz, whereas LP ones have been considered as the best choice for 8 and 10 kHz. However, there is no experimental evidence for this choice. In this work we attempt to solve this point by a direct comparison of system performances in an isolated word recognition task.

Another problem which has not been completely solved is the role of preemphasis in speech analysis for recognition. Therefore, all the parameters are computed with and without preemphasis of the speech signal. The bilinear transformation of the time sequence is also introduced for FFTCEP and LPCEP to emulate the perceptual frequency distortion represented by the Bark scale.

Section 1 presents a general view of the problem of speech recognition and outlines the importance of the extraction of discriminative acoustic features. The mathematical computation of all the parameters is described in Section 2. The methodology of the experimental evaluation and some numerical data about the computation of the parameters are presented in Sections 3 and 4. Section 5 shows the results of the experiments. The technical report ends with some concluding remarks and the list of references.

INDICE

LISTA DE FIGURAS

LISTA DE TABLAS

1-. Introducción.

El Reconocimiento Automático del Habla (RAH) constituye un problema básico en la comunicación hombre-máquina. Bajo esta denominación se consideran, en realidad, tareas de diferente complejidad: reconocimiento de palabras aisladas, *word spotting*, reconocimiento de discurso continuo, etc. afectadas por diversos factores, entre ellos el tamaño del vocabulario (reducido, medio o grande), la dependencia o independencia del locutor, la capacidad de adaptación a nuevos locutores, el tipo de gramática, etc. En los años setenta, parcialmente solucionadas algunas de estas tareas, se perfilaban dos aproximaciones al problema del reconocimiento de discurso continuo: los Modelos Estructurales Estocásticos y los Sistemas Basados en el Conocimiento. Demostrada en los primeros ochenta la ineficacia de éstos últimos, se invierte todo el esfuerzo en desarrollar sistemas capaces de extraer conocimiento de forma inductiva-probabilística, es decir, a partir de muestras. Se utiliza, a partir de entonces, una modelización acústica basada en Modelos Ocultos de Markov (*Hidden Markov Models*, HMM), discretos y continuos, y se optimizan los algoritmos de aprendizaje para entrenar los sistemas a partir de grandes bases de datos. Actualmente la investigación se concentra, por un lado, en mejorar el rendimiento de la modelización acústica (HMM Semicontinuos, Entrenamiento Discriminativo) y, por otro lado, en la integración de niveles de conocimiento superiores (Modelización del Lenguaje).

Los sistemas de RAH constan de varias etapas. Parten de una señal de voz, a la cual aplican técnicas de procesamiento de señal y reconocimiento de patrones para generar una cadena de unidades lingüísticas y a continuación, dependiendo de la aplicación, producir una interpretación semántica que, a su vez, genere una acción u otro tipo de representación de alto nivel. Estas etapas van encadenadas secuencialmente e implican sucesivos filtrados no recuperables, por lo que cualquier pérdida de información ocurrida en una de ellas afecta irreversiblemente a todas las siguientes.

Por otra parte, los algoritmos de reconocimiento de patrones requieren reducir drásticamente el volumen de datos de la señal de voz. Para ello deberá eliminarse toda información redundante o inútil, y mantener sólo información relevante, a ser posible mediante un número pequeño de parámetros. Esta discriminación de información y reducción del volumen de datos es lo que trataremos de efectuar, en primer término, mediante la parametrización.

La primera etapa de procesamiento consiste en la conversión A/D de la señal de voz (filtrado antialiasing, muestreo y cuantificación). De ella se obtiene una secuencia de números, tratable computacionalmente, que no contiene toda la información acústica de la señal de voz original, pero sí toda la información que nos interesa a efectos de reconocimiento.

El siguiente paso es lo que denominamos propiamente *parametrización*. La secuencia de números se divide en pequeños segmentos consecutivos y solapados, cada uno de los cuales se analiza por separado y produce un vector de parámetros o *vector característico*. Pueden aplicarse distintos tipos de análisis: energía, cruces por cero, banco de filtros, transformada de Fourier, predicción lineal, etc. Es posible generar un único vector reuniendo dos o más representaciones distintas. Se trata de almacenar el máximo de información en un espacio mínimo. En cualquier caso, como resultado se obtiene una nueva secuencia, esta vez de vectores característicos. Precisamente, determinar la mejor forma de dar ese primer paso, en el que a partir de una forma de onda se obtiene una secuencia equivalente de vectores característicos, es el objeto de este trabajo.

2-. Parametrizaciones.

Denominamos *parametrización* de una señal de voz a su caracterización mediante un vector de parámetros. No es necesario poder recuperar la señal original a partir de dichos parámetros. La cuestión no es obtener una representación de alta fidelidad, sino captar únicamente información relevante a efectos de percepción auditiva y reconocimiento del mensaje.

La identificación fonética de un segmento de voz depende básicamente de su envolvente espectral. De ahí que los parámetros que conforman un vector característico deban constituir una representación equivalente de dicha envolvente. De hecho, una forma de establecer el grado de fidelidad de una representación parámetrica consiste en medir la distancia euclídea entre la envolvente parametrizada y la envolvente original.

Así pues, es la envolvente del espectro instantáneo lo que tratamos de caracterizar. Dicha envolvente varía con relativa lentitud. Puede admitirse, aunque no es rigurosamente cierto salvo para segmentos vocálicos, que en tramos de aproximadamente 20 ms las componentes espectrales permanecen estacionarias. Ello permite ir analizando la señal a tramos y generar sucesivos vectores de parámetros que constituyen una representación equivalente a efectos de reconocimiento.

Para mejorar la representación espectral, debe aplicarse a los tramos de señal una ventana de análisis con un cierto solapamiento que permita suavizar la evolución de la envolvente y captar transiciones acústicas muy breves pero importantes, que de otra forma pasarían desapercibidas.

Considérese como ejemplo una Transformada Discreta de Fourier Dependiente del Tiempo (TDDFT). El oído humano realiza básicamente la misma operación de análisis, con una escala de frecuencias distorsionada, correspondiente a una mayor sensibilidad a bajas que a altas frecuencias (escala Bark de bandas críticas). La TDDFT consta de módulo y fase, pero en periodos de tiempo pequeños el oído no es sensible a la relación de fase entre las componentes espectrales, captando únicamente el módulo de la TDDFT, y siendo particularmente sensible a las frecuencias de resonancia (*formantes*) que aparecen en su envolvente. Cuestiones como el tipo de excitación (sorda/sonora) aplicada en cada instante por el aparato fonador, la frecuencia fundamental (en el caso de tramos sonoros) y otras características propias de cada locutor, afectan en menor grado a la percepción del mensaje y no serán tenidas en cuenta en este trabajo.

Un submuestreo-promediado en bandas críticas del módulo de la TDDFT constituye, de hecho, una representación paramétrica de la señal de voz. Pero el vector característico puede resultar demasiado grande e inmanejable por los algoritmos de entrenamiento y reconocimiento. Por esta razón, se han buscado representaciones más compactas, aplicando otras técnicas de análisis.

Tal es el caso de la predicción lineal (*Linear prediction*, LP), que se basa en la posibilidad de predecir el valor de una muestra a partir de p muestras anteriores, mediante suma ponderada por un conjunto de coeficientes que varían en el tiempo. Para ello, el fenómeno descrito por la secuencia de muestras debe corresponder a un filtro todo-polo. El aparato fonador responde aproximadamente a las características de un filtro todo-polo, salvo en el caso de los sonidos nasales. No obstante, la elección de un orden p lo bastante elevado asegura una buena caracterización de los formantes por medio del espectro LP. El número de coeficientes necesario depende de la frecuencia de muestreo. En codificación suele corresponder a un polo complejo conjugado (2 coeficientes) por cada kHz de

espectro, más un polo complejo conjugado adicional para caracterizar la forma del pulso glotal [Papamichalis, 87].

La representación paramétrica más utilizada en RAH se basa en una transformación homomórfica denominada *cepstrum* -transformada inversa del logaritmo del espectro. En dicha transformación sólo tienen relevancia los primeros coeficientes, que constituyen la respuesta impulso del tracto vocal, visto como un filtro digital. Los coeficientes cepstrales pueden utilizarse para caracterizar de forma más compacta un espectro LP. Un conjunto de 10 coeficientes cepstrales puede caracterizar una envolvente espectral de calidad equivalente a la que producen 14 o más coeficientes LP [Lee, 89]. Si a la transformación cepstral se añade una distorsión de la escala de frecuencias similar a la que sucede en el oído, mediante una transformación bilineal [Oppenheim, 72], los resultados en reconocimiento deben mejorar sustancialmente.

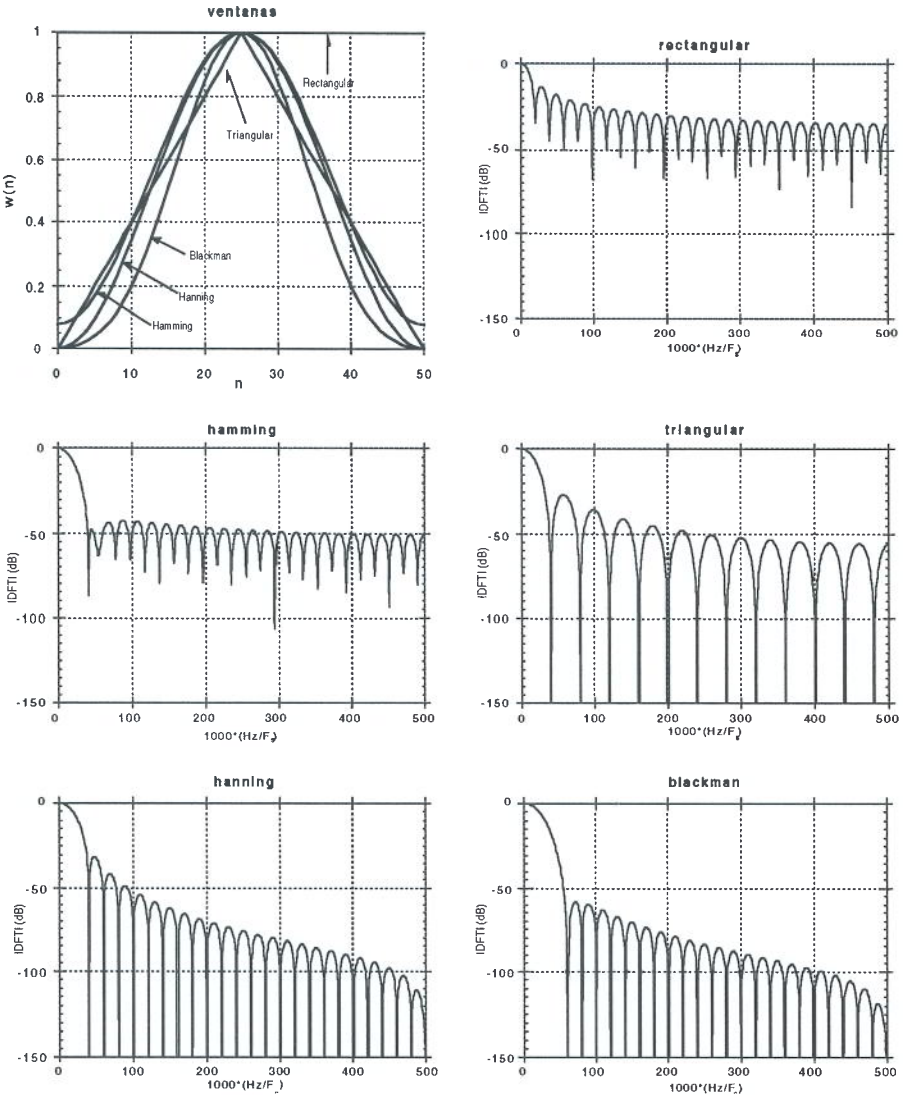

Figura 1. Forma y respuesta en frecuencia de algunas ventanas.

2.1-. Ventana.

Como se sabe, al aplicar una ventana a la señal de voz, la TDDFT que se obtiene es la convolución de la DFT del tramo de señal considerado y la DFT de la ventana aplicada. Una representación espectral de alta fidelidad requiere una ventana con respuesta en frecuencia de tipo impulsional (ancho de banda nulo y razón de filtrado infinita), es decir,

una ventana plana de longitud infinita, que en la práctica no puede aplicarse. Deberá diseñarse, pues, una aproximación finita a la ventana ideal, con las especificaciones de un filtro pasobajo, con ancho de banda mínimo y razón de filtrado máxima. Existen varias aproximaciones: ventanas rectangular, triangular, Hamming, Hanning, Blackman, etc. (Figura 1) [Rabiner, 78]. De todas ellas, la más utilizada, por su sencillez y eficacia, es la ventana Hamming:

$$w[n] = 0.54 - 0.46 \cdot \cos(2\pi n / (N-1)) \quad 0 \le n \le N-1 \tag{1}$$

La longitud de la ventana determina el tramo de señal en análisis. Al aumentar dicha longitud, disminuye el ancho de banda de la respuesta en frecuencia pero se mantiene la razón de filtrado. Por otra parte, si la ventana es demasiado pequeña, disminuye la resolución en frecuencia del análisis espectral, y análogamente, si la ventana es demasiado grande, disminuye la resolución en tiempo. Por tanto, la ventana debe ser suficientemente grande como para afinar el contenido espectral de la señal, pero también suficientemente pequeña como para no enmascarar movimientos espectrales contenidos en la misma. De esta forma, ambas características: la posición y la evolución de las componentes espectrales, deberán ser registradas con eficacia. Típicamente, la longitud de ventana se encuentra entre 20 y 40 ms.

Como se ha dicho, el solapamiento de ventanas permite captar y suavizar transiciones espectrales. Pero aumentar el solapamiento incrementa la frecuencia de submuestreo, es decir, el número de vectores por segundo, y ello repercute en el coste computacional de etapas posteriores del sistema de reconocimiento. La frecuencia de submuestreo suele estar comprendida entre 50 y 100 Hz.

2.2-. Preénfasis.

Un estudio detallado del aparato fonador [Rabiner, 78] revela que la onda de presión S(z) correspondiente a señales sonoras está afectada por tres procesos de filtrado. En primer lugar, la secuencia de excitación e[n], formada por un tren de pulsos, sufre un filtrado paso-bajo G(z) al atravesar la glotis. A continuación, la onda resultante atraviesa las sucesivas secciones del tracto vocal V(z), hasta llegar a los labios, donde presenta unas ciertas frecuencias de resonancia (formantes), que dependen de las citadas secciones. Finalmente, el flujo de aire producido en los labios es radiado al exterior como una onda de presión, según una función de transmisión R(z).

$$S(z) = R(z)V(z)G(z)E(z) \tag{2}$$

La glotis se modeliza mediante un filtro paso-bajo G(z) con un polo real doble. La frecuencia de corte de este filtro se sitúa alrededor de los 100 Hz. Se tiene:

$$G(z) = \frac{1}{\left(1 - e^{-cT} z^{-1}\right)^2} \quad , \tag{3}$$

donde c>0 es una constante y T el periodo de la excitación.

La función de transferencia del tracto vocal V(z) es un filtro todo-polo con un polo complejo conjugado por cada formante.

$$V(z) = \frac{1}{\prod_{i=1}^{K}\left[1 - 2e^{-c_i T}\cos(b_i T) z^{-1} + e^{-2c_i T} z^{-2}\right]} \quad , \tag{4}$$

donde c_i y b_i son constantes positivas que determinan el ancho de banda y la frecuencia central de cada formante, respectivamente.

El efecto de radiación R(z) -relación entre onda de presión y flujo de aire en los labios- corresponde, en primera aproximación, a un diferenciador:

$$R(z) = 1 - z^{-1} \tag{5}$$

Finalmente, la función de transferencia total H(z) -relación entre onda de presión y excitación- tiene la forma:

$$H(z) = \frac{S(z)}{E(z)} = \frac{1 - z^{-1}}{\left(1 - e^{-cT} z^{-1}\right)^2 \left\{ \prod_{i=1}^{K} \left[1 - 2e^{-c_i T} \cos(b_i T) z^{-1} + e^{-2c_i T} z^{-2} \right] \right\}} \tag{6}$$

Generalmente, cT<<1, lo cual permite aproximar G(z):

$$G(z) \cong \frac{1}{(1 - z^{-1})^2} = \frac{1}{[R(z)]^2} \tag{7}$$

De aquí obtenemos una nueva expresión para S(z):

$$S(z) = H(z)E(z) \cong \frac{V(z)}{R(z)} E(z) \tag{8}$$

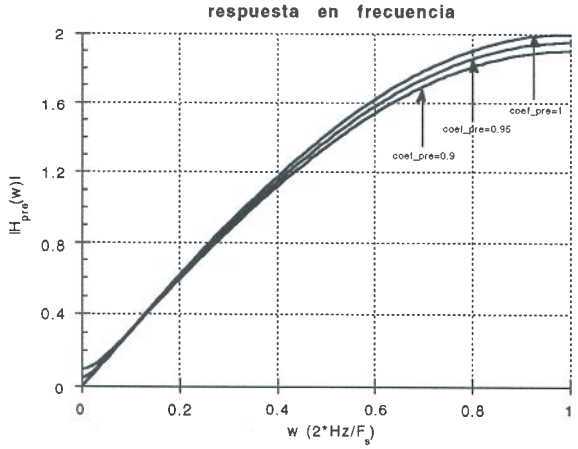

Figura 2. Respuesta en frecuencia de varios filtros de preénfasis, con coeficientes 0.9, 0.95 y 1, respectivamente. Suele utilizarse un coeficiente menor que la unidad para preservar la estabilidad del filtro.

Si el modelo en el que se basa nuestra representación paramétrica trata de describir la configuración articulatoria V(z), es decir, la secuencia de secciones principales del tracto vocal, visto como un tubo acústico, interesa eliminar de la onda de presión las aportaciones debidas a filtrado glotal y radiación labial, para quedarnos únicamente con la aportación debida al tracto vocal. A tal efecto, en una etapa previa al análisis, se aplica a la señal la siguiente función de transferencia:

$$H_{pre}(z) = 1 - az^{-1} \qquad a <\cong 1 \tag{9}$$

denominada *filtro de preénfasis*, ya que básicamente se trata de un diferenciador, tipo R(z), que enfatiza linealmente las componentes de alta frente a las de baja frecuencia. La respuesta en frecuencia, tal como se aprecia en la Figura 2, tiene aspecto de rampa. Para obtenerla analíticamente, sólo tenemos que evaluar la transformada z del filtro de preénfasis en el círculo unidad:

$$H_{pre}(e^{j\omega}) = 1 - ae^{-j\omega} \tag{10}$$

$$\left| H_{pre}(e^{j\omega}) \right|^2 = 1 + a^2 - 2a\cos(\omega) \tag{11}$$

$$a = 1 \Rightarrow \text{diferenciador: } \left|H_{pre}(e^{j\omega})\right|^2 = 4sen^2(\omega/2) \qquad (12)$$

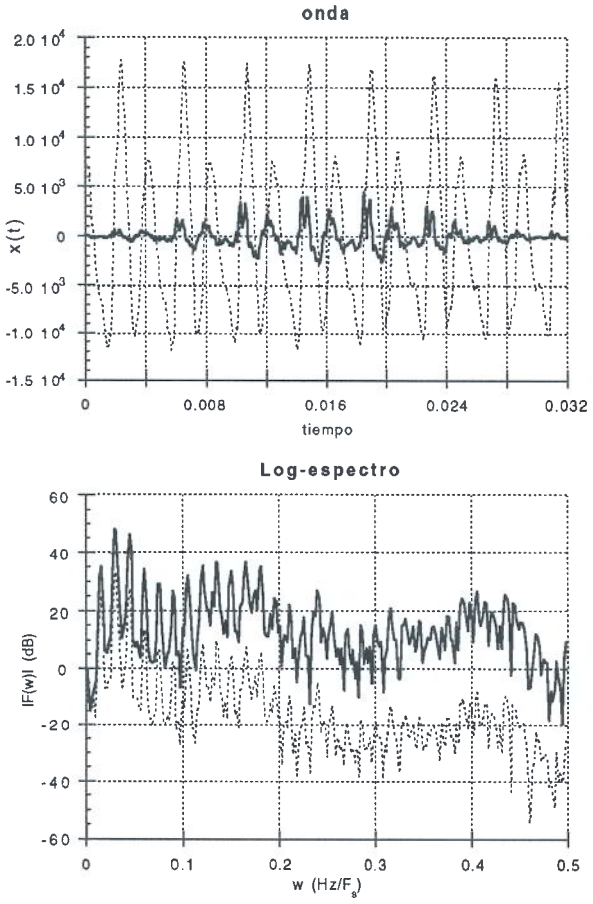

Figura 3. La primera gráfica muestra una onda a la que se ha aplicado una ventana Hamming y preénfasis (coeficiente: 0.95), junto a la onda original (con trazo punteado). Debajo se muestra el log-espectro de ambas, observándose una elevación característica de las componentes de alta frecuencia.

Cuando la señal de voz corresponde a un segmento sordo, no hay razón para aplicar preénfasis. El mecanismo de excitación es distinto, sin intervención de la glotis. La excitación se produce en un punto intermedio del tracto vocal, y por tanto sólo actúan las últimas secciones del mismo, de forma que es necesario un menor número de polos para modelizar las resonancias. El efecto de radiación labial continúa enfatizando las componentes de alta frecuencia. Aplicar preénfasis supondría duplicar dicho efecto. En todo caso, podríamos aplicar un filtro inverso de preénfasis, con objeto de compensar el efecto de radiación y recuperar la onda emitida en los labios.

Por otra parte, tampoco tiene sentido aplicar preénfasis cuando la representación paramétrica se basa en una modelización del aparato auditivo, como sucede con la transformada de Fourier y con el banco de filtros. No obstante, conviene comprobar el rendimiento de todas las representaciones paramétricas con y sin filtro de preénfasis.

2.3-. Banco de filtros con escala Bark.

La Transformada Discreta de Fourier (DFT) de una secuencia x[n] es una secuencia compleja definida por la siguiente expresión:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi nk/N} \quad k = 0..N-1 \qquad (13)$$

Si secuencia y transformada tienen la misma longitud N, X[k] constituye una representación exacta de x[n]. De hecho, se define la transformada inversa:

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]\cdot e^{j2\pi nk/N} \quad n = 0..N-1 \tag{14}$$

Claramente, si secuencia y transformada tienen la misma longitud, uno de los objetivos fundamentales de la parametrización, a saber, la reducción del volumen de datos, no se consigue. Como se ha dicho más arriba, sólo nos interesa el módulo de la DFT, ya que en tramos de señal pequeños el oído no es capaz de percibir la relación de fase entre las componentes espectrales. Por otra parte, la transformada contiene información relativa a la periodicidad de la señal y a características propias del locutor, lo cual incrementa el tamaño de la representación pero no es útil a efectos de reconocimiento.

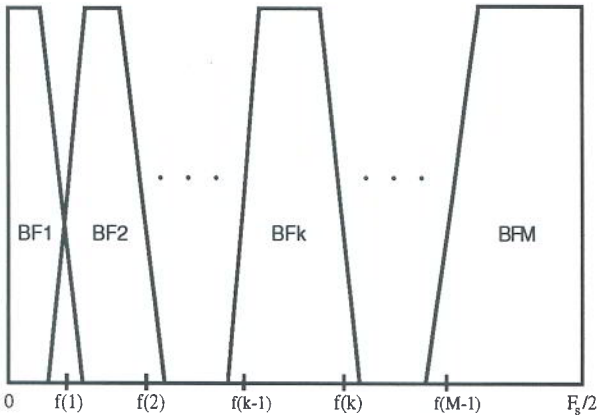

Figura 4. Promediado en bandas de una DFT. El módulo de la DFT es ponderado mediante una secuencia de ventanas trapezoidales. Posteriormente se obtiene un promedio aritmético de la energía espectral en cada banda de frecuencia.

En definitiva, se trata de suavizar el perfil del módulo de la DFT para obtener su envolvente. Existen varias formas de aproximar dicha envolvente. Una de las más sencillas consiste en realizar un submuestreo en frecuencia, mediante promediado en bandas de los coeficientes de la transformada de Fourier (Figura 4). En cada banda de frecuencia, tal como muestra la ecuación (15), se aplica una ventana trapezoidal que pondera los coeficientes de la DFT antes de efectuar el promediado.

Esta idea es equivalente a la de aplicar a la señal un banco de filtros pasabanda (BF). En la práctica, un banco de filtros suele realizarse mediante modulación a banda base y filtrado pasobajo, obteniendo un conjunto de coeficientes que representan la energía de la señal en cada una de las bandas de frecuencia consideradas. Pueden utilizarse filtros equidistantes de ancho de banda constante. Esto significa suponer que la resolución en frecuencia del oído humano es constante. No obstante, ciertos experimentos psicoacústicos [Zwicker, 81] han revelado la existencia de bandas críticas (Tabla I), bandas naturales de percepción cuya anchura va aumentando con la frecuencia. Una parametrización basada en un modelo de percepción auditiva debe tener en cuenta este fenómeno.

$$BF_{DFT}[k]=\begin{cases}\sum_{i=0}^{f_1-1}|X[i]|^2+\frac{1}{2}|X[f_1]|^2 & k=1\\ \frac{1}{2}|X[f_{k-1}]|^2+\sum_{i=f_{k-1}+1}^{f_k-1}|X[i]|^2+\frac{1}{2}|X[f_k]|^2 & 1<k<M\\ \frac{1}{2}|X[f_{M-1}]|^2+\sum_{i=f_{M-1}+1}^{N-1}|X[i]|^2 & k=M\end{cases} \quad (15)$$

Tabla I. Bandas críticas detectadas en el oído humano. Se indican la frecuencia central (f_m), el ancho de banda (Δf) y el límite superior (F_{max}) de cada banda de frecuencia. El límite inferior de la primera banda se sitúa en 20 Hz.

N	f_m (Hz)	Δf (Hz)	10 log(Δf/Hz) (dB)	F_{max} (Hz)
1	50	80	19	100
2	150	100	20	200
3	250	100	20	300
4	350	100	20	400
5	450	110	20	510
6	570	120	21	630
7	700	140	21	770
8	840	150	22	920
9	1000	160	22	1080
10	1170	190	23	1270
11	1370	210	23	1480
12	1600	240	24	1720
13	1850	280	25	2000
14	2150	320	25	2320
15	2500	380	26	2700
16	2900	450	27	3150
17	3400	550	27	3700
18	4000	700	28	4400
19	4800	900	29	5300
20	5800	1100	30	6400
21	7000	1300	32	7700
22	8500	1800	32	9500
23	10500	2500	34	12000
24	13500	3500	35	15500

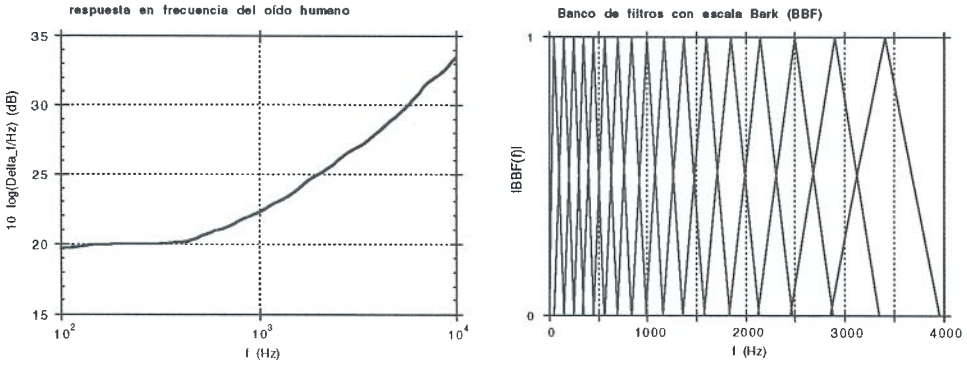

Figura 5. Respuesta en frecuencia del oído humano y banco de filtros (hasta 4000 Hz), correspondientes a las bandas críticas de la Tabla I.

Aplicar un banco de filtros con bandas críticas es equivalente a considerar una escala de frecuencias distorsionada, denominada *escala Bark*, que viene dada aproximadamente, cuando la frecuencia de muestreo es 16 kHz, por la siguiente expresión:

$$F_{bark}=1.95\ arctg\left(\frac{F_{Hz}}{1316}\right)+0.525\ arctg\left(\frac{F_{Hz}}{7500}\right)^2 \quad (16)$$

Si se admite como hipótesis que el habla ha evolucionado para adaptarse a las características del aparato auditivo, la utilización de bandas críticas debe contribuir a

aumentar las tasas de reconocimiento, ya que supone una importante mejora en la modelización.

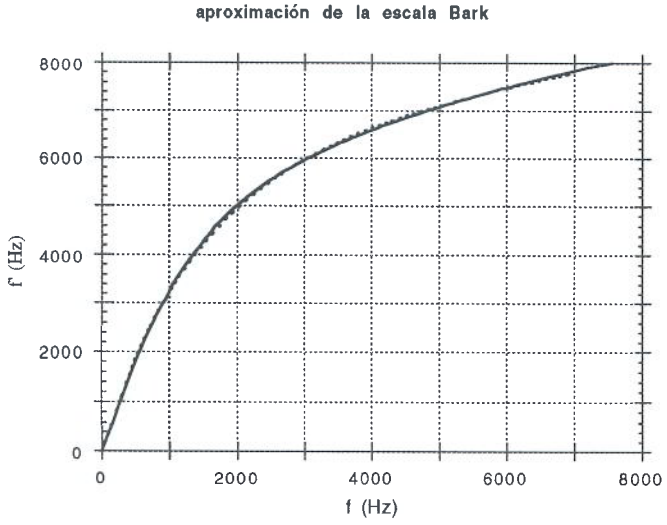

Figura 6. Transformación de la escala de frecuencias según la escala Bark de bandas críticas (línea punteada), y aproximación analítica expresada en la fórmula (16) (trazo continuo).

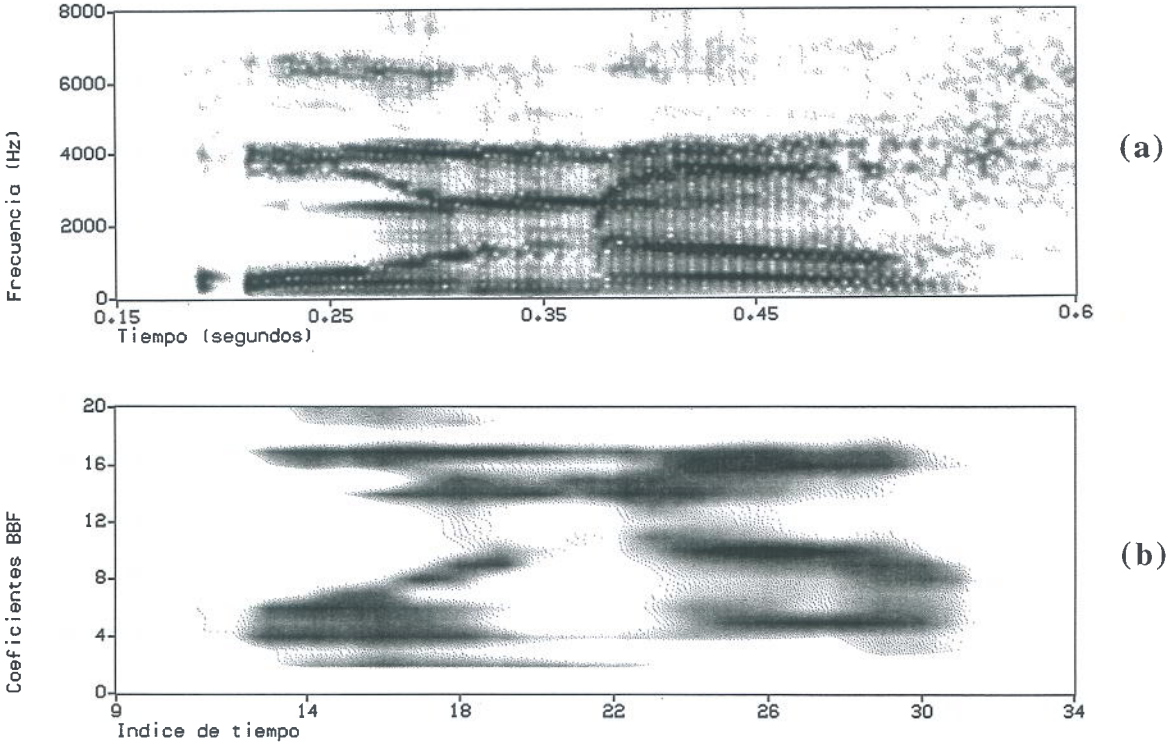

Figura 7. Espectrograma original (a) y promediado en bandas de frecuencia con escala Bark y preénfasis (b) correspondientes a una pronunciación de la palabra /uno/. Puede observarse una ampliación del espectro en la región de baja frecuencia con respecto a la región de alta frecuencia.

2.4-. Coeficientes cepstrales a partir de DFT.

La señal de voz s[n] puede obtenerse mediante la convolución en el tiempo de dos señales elementales: la respuesta impulso de un filtro digital $v_s[n]$ y una secuencia p[n] que actúa como mecanismo de excitación.

$$s[n] = \sum_{k=-\infty}^{+\infty} v_\delta[k]p[n-k] \tag{17}$$

$$S(z) = V(z)P(z) \tag{18}$$

Los coeficientes del filtro V(z) modelizan la articulación instantánea del tracto vocal, es decir, determinan la posición de los formantes. Por esta razón, constituyen una representación paramétrica adecuada. Por otra parte, la respuesta impulso $v_\delta[n]$, en general de longitud infinita, constituye una representación equivalente del filtro.

$$V(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{r=0}^{M} b_r z^{-r}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{19}$$

$$x[n] = \delta[n] \;\Rightarrow\; v_\delta[n] = \sum_{k=1}^{N} a_k v_\delta[n-k] + \sum_{r=0}^{M} b_r \delta[n-r] \tag{20}$$

Al invertir la convolución podemos conocer la respuesta impulso del tracto vocal $v_\delta[n]$, que, de esta forma, podría utilizarse como representación paramétrica, truncándola según convenga. Esto es precisamente lo que se consigue mediante una transformación denominada *cepstrum* [Rabiner, 78] [Oppenheim, 89]. El *cepstrum complejo* viene dado por la siguiente expresión:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\omega}) e^{j\omega n} d\omega \quad , \tag{21}$$

donde

$$\hat{X}(e^{j\omega}) = \log\left\{X(e^{j\omega})\right\} = \log\left|X(e^{j\omega})\right| + j\arg\left[X(e^{j\omega})\right] \tag{22}$$

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n] e^{-j\omega n} \tag{23}$$

Esta transformación convierte la convolución en suma:

$$\begin{aligned} \hat{S}(e^{j\omega}) &= \log\left\{V(e^{j\omega})P(e^{j\omega})\right\} = \\ &= \log\left\{V(e^{j\omega})\right\} + \log\left\{P(e^{j\omega})\right\} = \\ &= \hat{V}(e^{j\omega}) + \hat{P}(e^{j\omega}) \end{aligned} \tag{24}$$

$$\hat{s}[n] = \hat{v}_\delta[n] + \hat{p}[n] \tag{25}$$

La secuencia c[n] denominada simplemente *cepstrum* se define:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log\left|X(e^{j\omega})\right| e^{j\omega n} d\omega \tag{26}$$

Esta secuencia corresponde a la parte simétrica conjugada del cepstrum complejo. El módulo de la transformada de Fourier de una secuencia real es una función par. Por tanto, los coeficientes c[n] constituyen a su vez una secuencia real, la cual puede obtenerse mediante la siguiente expresión:

$$c[n] = \frac{1}{\pi} \int_{0}^{\pi} \log\left|X(e^{j\omega})\right| \cos(\omega n) d\omega \tag{27}$$

En la práctica los coeficientes cepstrales se calculan de forma aproximada utilizando transformadas discretas de Fourier:

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi}{N}kn} \quad k=0..N-1 \tag{28}$$

$$c[n]=\frac{1}{N}\sum_{k=0}^{N-1}\log|X[k]|e^{j\frac{2\pi}{N}kn} \quad n=0..N-1 \tag{29}$$

2.5-. Coeficientes cepstrales a partir de BF.

Como sabemos, los coeficientes de un banco de filtros determinan una versión suavizada del espectro -una especie de envolvente. Si se aplica una transformada de cosenos al logaritmo de dichos coeficientes, se obtiene un conjunto de coeficientes cepstrales que caracterizan la citada envolvente. De alguna forma, estos coeficientes han de tener mayor calidad que los obtenidos a partir de una DFT, ya que incorporan una menor cantidad de información redundante. Se tiene, pues:

$$c[n]=\sum_{k=1}^{M}\log(BF[k])\cos[(2k-1)\pi n/2M] \quad n=1..C\ , \tag{30}$$

donde M es el número de filtros, y C el número de coeficientes cepstrales.

La utilización de bandas críticas introduce una distorsión de la envolvente similar a la que tiene lugar en el oído interno. Ello permite reducir el número de bandas de frecuencia, y posteriormente, una vez realizada la transformación cepstral, reducir también el tamaño del vector característico, obteniendo una representación más compacta.

2.6-. Análisis de predicción lineal.

El análisis de predicción lineal (*Linear Prediction*, LP) utiliza un filtro digital todo-polo para modelizar dinámicamente la función de transferencia del tracto vocal, mediante una minimización del error de predicción en tramos sucesivos de la señal de voz. Dicha minimización conduce a un sistema de ecuaciones lineales de fácil resolución, cuya matriz característica, dependiendo del método concreto de análisis, permite aplicar algoritmos de cálculo muy eficientes. La modelización debe ser tanto mejor cuanto mayor sea el número de coeficientes del filtro. Se estima necesario establecer un polo complejo conjugado por cada kHz de señal. Esto significa que si aumentamos la frecuencia de muestreo necesitaremos un número proporcionalmente mayor de coeficientes.

En primer lugar, se obtiene una estimación de la señal a partir de una combinación lineal de p muestras anteriores:

$$\hat{s}[n]=\sum_{k=1}^{p}a_k s[n-k] \tag{31}$$

El error de predicción viene dado por el filtro $A(z)$:

$$A(z)=1-\sum_{k=1}^{p}a_k z^{-k} \tag{32}$$

$$e[n]=s[n]-\sum_{k=1}^{p}a_k s[n-k] \tag{33}$$

La señal es recuperada a partir de los coeficientes de predicción lineal, los cuales configuran un filtro digital todo-polo *H(z)*, atacado por una secuencia de excitación *u[n]*, que en el caso ideal debe coincidir con el error de predicción *e[n]*.

$$H(z)=\frac{S(z)}{U(z)}=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{34}$$

$$\bar{s}[n]=Gu[n]+\sum_{k=1}^{p}a_k\bar{s}[n-k] \tag{35}$$

Los coeficientes óptimos se obtienen minimizando el error cuadrático medio de predicción en un tramo de señal:

$$E_n=\frac{1}{N}\sum_m\left(s_n[m]-\sum_{k=1}^{p}a_k s_n[m-k]\right)^2 \quad \text{MINIMO,} \tag{36}$$

donde $s_n[m]=s[n+m]$ es un segmento de señal alrededor de la muestra n, y N la longitud de dicho segmento.

Para minimizar E_n, deben ser nulas todas sus derivadas con respecto a los a_i:

$$\frac{\partial E_n}{\partial a_i}=0 \quad 1\le i\le p \tag{37}$$

De aquí resulta el siguiente sistema de ecuaciones lineales:

$$\sum_{k=1}^{p}a_k\phi_n(i,k)=\phi_n(i,0) \quad 1\le i\le p, \tag{38}$$

donde

$$\phi_n(i,k)=\sum_m s_n[m-i]s_n[m-k] \quad 1\le i\le p \quad 0\le k\le p \tag{39}$$

Para resolver este sistema de ecuaciones de forma computacionalmente eficiente, se han propuesto dos métodos [Rabiner, 78]: autocorrelación y covarianza. Básicamente se distinguen en la definición del intervalo de suma sobre el que se evalua el error cuadrático medio de predicción.

El método de autocorrelación analiza una señal infinita, nula fuera del intervalo considerado. Para suavizar la señal en los extremos del intervalo se le aplica una ventana. A continuación se obtiene el error de predicción total, y posteriormente los coeficientes de autocorrelación, dando lugar a una matriz característica simétrica de Toeplitz, para cuya resolución se aplica un algoritmo recursivo (Durbin) que asegura soluciones estables [Makhoul, 75].

$$s_n[m]=s[m+n] \quad m=0..N-1 \tag{40}$$

$$E_n=\sum_{m=0}^{N-1}e_n^2[m] \tag{41}$$

$$\phi_n(i,k)=R_n(|i-k|) \quad i=1..p \quad k=0..p \qquad \text{MATRIZ TOEPLITZ,} \tag{42}$$

donde

$$R_n(j)=\sum_{m=0}^{N-1-j}s_n[m]s_n[m+j] \quad j=0..p \tag{43}$$

Algoritmo de Durbin

$$E_n^{(0)} = R_n(0)$$

$$k_i = \left(R_n(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R_n(i-j) \right) / E_n^{(i-1)} \qquad i = 1..p$$

$$a_i^{(i)} = k_i$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad j = 1..i-1$$

$$E_n^{(i)} = (1 - k_i^2) E_n^{(i-1)}$$

$$a_j = a_j^{(p)} \qquad j = 1..p$$

Figura 8. Algoritmo recursivo de Durbin para la obtención de los coeficientes LP por el método de autocorrelación. Los coeficientes intermedios k_i se denominan *coeficientes de reflexión.*

El método de covarianza no requiere ventana. Se calcula el error de predicción en el intervalo considerado, manteniendo la señal inalterada.

$$s_n[m] = s[m+n] \qquad m = 0..N-1 \tag{44}$$

$$E_n = \sum_{m=0}^{N-1} e_n^2[m] \tag{45}$$

En la matriz característica, simétrica no de Toeplitz, aparecen correlaciones cruzadas de dos tramos finitos próximos pero distintos:

$$\phi_n(i,k) = \phi_n(k,i) = \sum_{m=0}^{N-1} s_n[m-i] s_n[m-k] \qquad i = 1..p \quad k = 0..p \tag{46}$$

Fácilmente se puede comprobar que:

$$\begin{aligned} \phi_n(i+1,k+1) = \phi_n(i,k) \\ + s_n[-i-1] s_n[-k-1] \\ - s_n[N-i-1] s_n[N-k-1] \end{aligned} \qquad \forall i,k \tag{47}$$

por lo que la matriz característica, aunque simétrica, no es Toeplitz. El algoritmo de resolución es menos eficiente y no asegura soluciones estables [Markel, 76].

La configuración computacional más eficiente, equivalente al método de autocorrelación, se basa en una red en la que dos secuencias, error de predicción hacia adelante *e[n]* y error de predicción hacia atrás *b[n]*, definidas a partir de la fórmula recursiva de Durbin, y circulando en sentidos contrarios, interactúan a través de los denominados coeficientes de reflexión (*Reflection Coefficients*, RC). De alguna forma, esta modelización corresponde a una interpretación exacta de lo que sucede en el tracto vocal con las ondas incidente y reflejada. Partiendo directamente de la señal de voz, sin ventana, se generan de forma recursiva las secuencias de error hacia adelante y hacia atrás y los coeficientes de reflexión, también llamados coeficientes PARCOR, ya que se calculan a partir de la correlación parcial normalizada de *e[n]* y *b[n]* [Gómez, 87]. Esta aproximación resuelve el problema en un único paso y no requiere el cálculo explícito de la matriz de autocorrelación ni de los coeficientes de predicción.

Se definen las secuencias de error hacia adelante *e[n]* y de error hacia atrás *b[n]*:

$$\begin{cases} e_n^{(i)}[m] = s_n[m] - \sum_{k=1}^{i} a_k^{(i)} s_n[m-k] \\ b_n^{(i)}[m] = s_n[m-i] - \sum_{k=1}^{i} a_k^{(i)} s_n[m-i+k] \end{cases} \tag{48}$$

Desarrollando la fórmula recursiva de Durbin, se demuestran las siguientes relaciones:

$$\begin{cases} e_n^{(i)}[m] = e_n^{(i-1)}[m] - k_i b_n^{(i-1)}[m-1] \\ b_n^{(i)}[m] = b_n^{(i-1)}[m-1] - k_i e_n^{(i-1)}[m] \end{cases} \tag{49}$$

cuyas condiciones iniciales son:

$$e_n^{(0)}[m] = b_n^{(0)}[m] = s_n[m] \tag{50}$$

Finalmente, los coeficientes de reflexión pueden obtenerse a partir de las secuencias *e[n]* y *b[n]*. Existen varias aproximaciones (Itakura, Burg, etc.). Concretamente, la fórmula de Itakura [Rabiner, 78] se basa en la minimización conjunta del error cuadrático de predicción hacia adelante y hacia atrás:

$$k_i = \frac{\sum_{m=0}^{N-1} e_n^{(i-1)}[m] b_n^{(i-1)}[m-1]}{\left(\sum_{m=0}^{N-1} \left(e_n^{(i-1)}[m] \right)^2 \sum_{m=0}^{N-1} \left(b_n^{(i-1)}[m-1] \right)^2 \right)^{1/2}} \quad i = 1..p \tag{51}$$

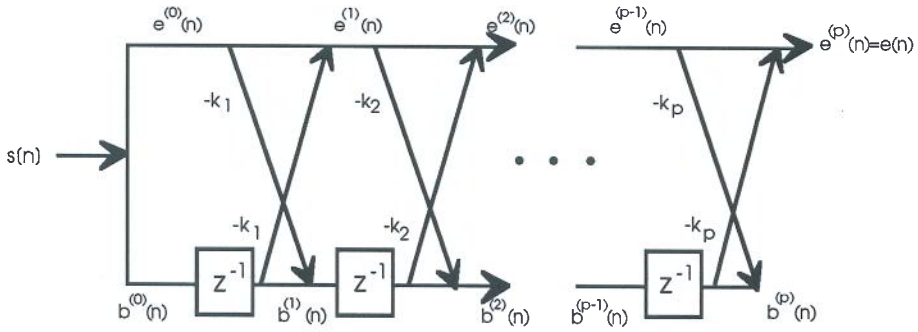

Figura 9. Estructura de cálculo de una red PARCOR.

El rango de variación de los coeficientes de predicción lineal es demasiado grande, por lo que pequeños errores de cuantificación pueden producir un filtro inestable. En su lugar se utilizan los coeficientes de reflexión, ya que basta con que todos ellos estén comprendidos entre -1 y 1 para asegurar la estabilidad del filtro de predicción.

Sin embargo, no parece conveniente utilizar una distancia euclídea directamente sobre los coeficientes de reflexión, ya que su distribución estadística no es uniforme. Para compensar en parte este efecto, se define, a partir de los RC, otro conjunto de coeficientes, los denominados *Log-Area Ratios* (LAR) [Rabiner, 78], cuya distribución estadística es relativamente uniforme:

$$LAR_i = \log\left(\frac{1+k_i}{1-k_i} \right) \quad i = 1..p \tag{52}$$

2.7-. Coeficientes cepstrales a partir de LP.

El espectro LP, aproximación de la envolvente del espectro de la señal, se obtiene evaluando *H(z)* en el círculo unidad.

$$H(e^{j\omega}) = \frac{S(e^{j\omega})}{U(e^{j\omega})} = \frac{G}{1 - \sum_{k=1}^{p} a_k e^{-j\omega k}} \tag{53}$$

Una representación equivalente, basada en coeficientes cepstrales, puede obtenerse mediante transformada inversa de Fourier del logaritmo de $H(e^{j\omega})$, aunque suele aplicarse la siguiente fórmula recursiva [Rabiner, 78]:

$$c[n] = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c[k] a_{n-k} \quad 1 \le n \le p \tag{54}$$

2.8-. Transformación Bilineal.

No siempre puede aplicarse directamente una distorsión de la escala de frecuencias equivalente a la introducida por el oído humano. Lo más inmediato parece aplicar en el espacio de la frecuencia la expresión analítica (16) que define dicha distorsión, pero esto no es posible cuando se manejan espacios de representación distintos a la frecuencia.

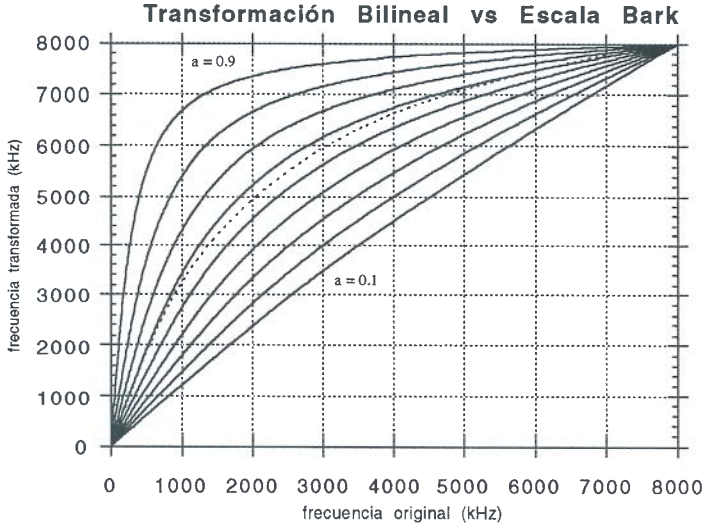

Figura 10. Relación entre frecuencia original y frecuencia transformada, para distintos valores del parámetro a de la transformación bilineal (desde a=0.1 hasta a=0.9), y escala Bark (línea punteada).

Como alternativa, se plantea una transformación de la señal en el espacio del tiempo que produzca, de forma indirecta, la misma distorsión de la escala de frecuencias que estamos buscando. En concreto, suele aplicarse la denominada transformación bilineal (*Bilinear Transform*, BLT) [Oppenheim, 72][Lee, 89], definida por la siguiente expresión:

$$\hat{z} = \frac{1 - az^{-1}}{z^{-1} - a} \tag{55}$$

Esta transformación implica simplemente un cambio de variable en z. El parámetro *a* determina las características de dicho cambio de variable (Fig. 10). La relación entre

frecuencia original (ω) y frecuencia transformada (Ω), se obtiene evaluando la transformación en el círculo unidad:

$$\Omega = \omega + 2arctg\left[\frac{a\sin\omega}{1-a\cos\omega}\right] \tag{56}$$

Una distorsión de la escala de frecuencias casi equivalente a la escala Bark puede obtenerse haciendo a=0.56 (Fig. 11).

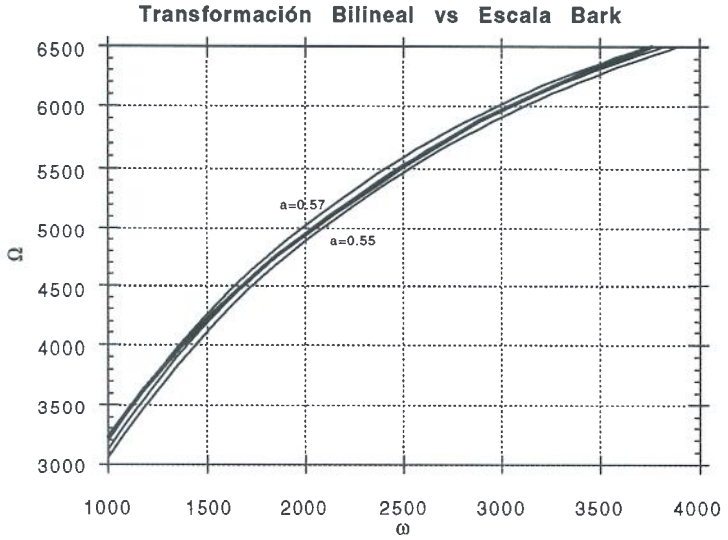

Figura 11. Transformación bilineal de la escala de frecuencias (trazos finos, para a=0.55, a=0.56 y a=0.57) y escala Bark (trazo grueso). Puede observarse cómo con a=0.56 se obtiene una buena aproximación de la escala de percepción auditiva.

La transformación bilineal puede aplicarse a la señal antes de proceder al análisis, o a los coeficientes cepstrales como última etapa del análisis. Se aplica, en cualquier caso, a una secuencia en el tiempo, obteniendo una nueva secuencia correspondiente a la variable z transformada. La implementación más eficaz consiste en una red de filtros digitales en cascada [Oppenheim, 72], que permite obtener directamente la secuencia transformada $g[k]=g_0[k]$ a partir de la secuencia original f[n] (Fig. 12).

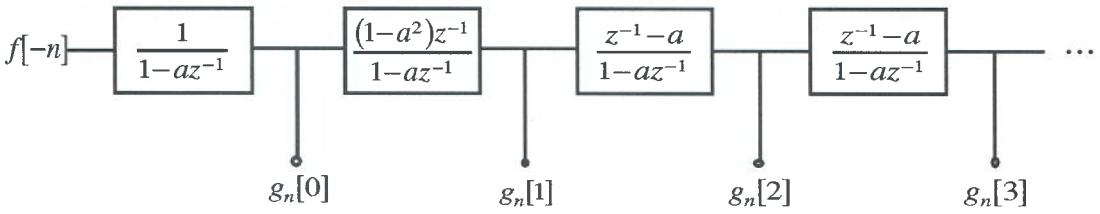

Figura 12. Implementación de la transformación bilineal mediante una red de filtros digitales. La secuencia transformada $g[k]=g_0[k]$ se genera directamente a partir de la secuencia original f[n].

Como puede verse, si se exceptúan los dos primeros bloques de la cadena, se tiene una estructura recursiva, de fácil implementación por software. Esta es una de las ventajas adicionales de la transformación bilineal frente a la utilización explícita de la escala Bark [Lee, 89]. Sin embargo, la BLT puede introducir correlaciones entre las componentes del vector de parámetros, lo cual invalidaría ciertas hipótesis (matriz de correlación diagonal) utilizadas por algunas metodologías de RAH [Bellegarda, 90] [Huang, 93].

2.9-. Energía.

La energía de la señal puede incorporarse como dimensión adicional en el vector de parámetros, ya que contiene información relevante desde el punto de vista del reconocimiento. De hecho, puede utilizarse para distinguir de forma bastante primitiva segmentos sonoros (vocales) de segmentos sordos (silencios, la mayor parte de las consonantes) .

Se define la energía E_m de un segmento de señal s_m:

$$E_m = \sum_{n=0}^{N-1} s_m^2[n] \quad , \tag{57}$$

donde N es la longitud del segmento.

Diversos experimentos han demostrado que la variación de la energía aporta más información que la propia energía, ya que determina, en muchos casos, el carácter estacionario o transitorio del segmento de señal[1] [Segura, 91]. Se define dicha variación:

$$\Delta E_m = E_m - E_{m-1} \tag{58}$$

Una distancia euclídea requiere, en teoría, rangos de variación iguales para todas las componentes del vector de parámetros. La incorporación de la energía o de su derivada al vector de parámetros exige ciertos ajustes, ya que normalmente el rango de variación de estas cantidades es varios órdenes de magnitud superior al rango de variación del resto de los parámetros. Suele aplicarse, en primer lugar, una transformación logarítmica, que proporciona rangos de variación del mismo orden de magnitud, y a continuación un reescalado estadístico, basado en las varianzas de los parámetros acústicos. Este reescalado puede hacer la varianza de la energía igual a la varianza media o a la varianza máxima del vector de parámetros, o igual a una extrapolación de las varianzas de las componentes de dicho vector. El peso estadístico asignado a la energía puede establecerse experimentalmente con el criterio de optimizar las tasas de reconocimiento [Segura, 91].

2.10-. *Liftering*.

Todas las representaciones paramétricas consideradas en este capítulo son, en el caso continuo, de longitud infinita, con lo cual la caracterización exacta de un segmento de señal exigiría un número infinito de parámetros. En la práctica se utilizan representaciones discretas cuya longitud N, aunque finita, continua siendo demasiado elevada. El tamaño del vector característico puede reducirse cuando, como en el caso de los coeficientes cepstrales, las varianzas decrecen con el orden del parámetro, es decir:

$$\sigma^2[n] < \sigma^2[n-1] \quad 1 < n \leq N \tag{59}$$

Esto significa que el espacio vectorial de los parámetros acústicos no es homogéneo y que las sucesivas componentes del vector de parámetros tendrían contribuciones estadísticamente decrecientes en una distancia euclídea. Esta propiedad permite despreciar aquellos parámetros cuya varianza sea inferior a un cierto umbral, pues se considera que no aportan cantidades significativas a la distancia. La representación quedaría truncada, y tendría una longitud L<N. Finalmente, antes de calcular la distancia, el decrecimiento de las varianzas es compensado mediante un cierto factor lineal $l(n)$ que enfatiza las componentes de orden alto. Nos queda un vector de longitud reducida cuyas componentes tienen varianzas similares.

---

[1] En general, las características dinámicas aportan más información que las características estáticas. Al parecer, ello es debido a que la información psicoacústica se encuentra en las transiciones espectrales.

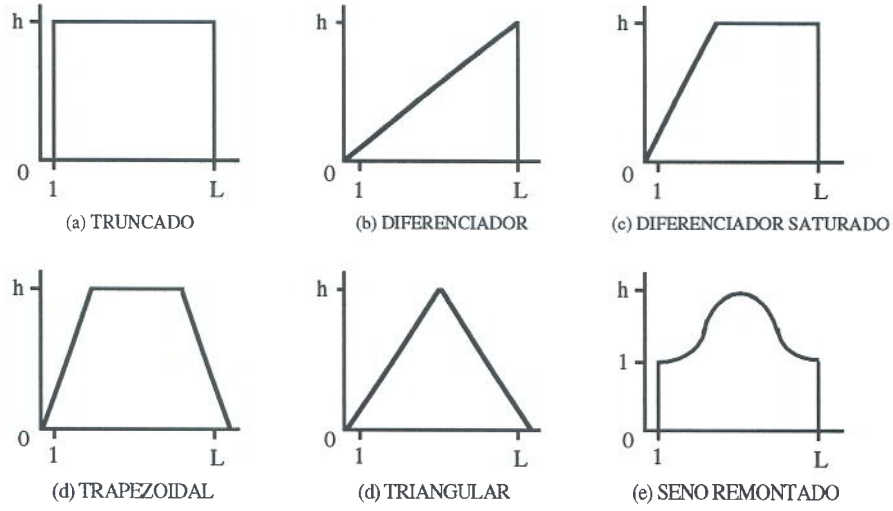

Figura 13. Ventanas de *liftering*.

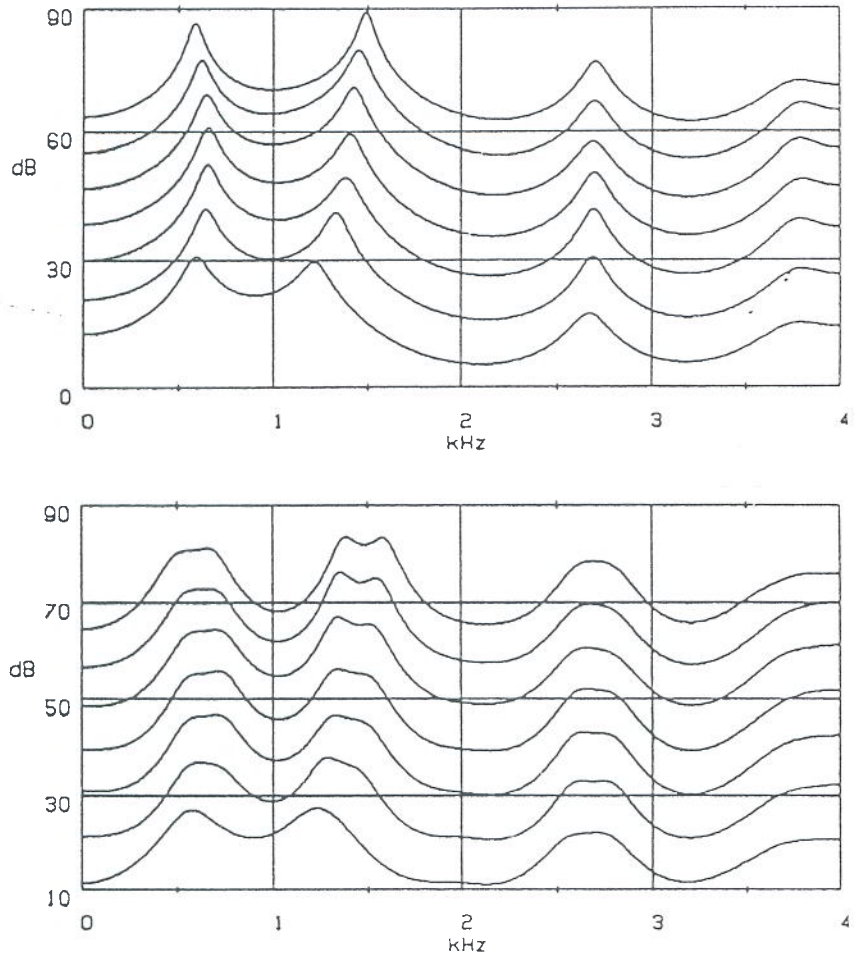

Figura 14. Espectros LP originales (a) y filtrados con seno remontado (b) [Segura, 91].

Este factor de compensación lineal $l(n)$ puede generalizarse mediante lo que denominamos ventana de *liftering*. Se trata de una ventana de longitud L<N que pondera

las componentes del vector característico. El efecto de esta ventana es enfatizar o deenfatizar determinadas componentes del vector de parámetros. Se han propuesto varias ventanas de *liftering* (Fig. 13), que han sido aplicadas específicamente sobre los coeficientes cepstrales [Tohkura, 87]. Los mejores resultados se obtienen con la ventana tipo seno remontado [Segura, 91]:

$$l[n] = 1 + \frac{L}{2}\sin\left(\frac{n\pi}{L}\right) \quad 1 \le n \le L \tag{60}$$

Ello es debido al deénfasis conjunto de los coeficientes cepstrales de orden bajo y de orden alto, que presentan variabilidades no deseadas para procesos de comparación de patrones [Tohkura, 87]. En concreto, los coeficientes de orden bajo presentan una variabilidad provocada por las condiciones del canal de transmisión utilizado en la adquisición de la señal y por características propias del locutor, como la forma del pulso glotal. Los coeficientes de orden alto deben su variabilidad al proceso de análisis.

La ventana de *liftering* puede verse como un filtro en el espacio de parámetros que, en el caso de los coeficientes cepstrales, produce un allanamiento del espectro y un ensanchamiento de los picos correspondientes a los formantes, lo cual reduce la sensibilidad de la distancia a la posición de dichos picos, manteniendo las características globales del espectro (Fig. 14) [Segura, 91]. En realidad, la modificación del peso estadístico relativo de los parámetros acústicos tiene que ver con la definición de una nueva distancia no euclídea. Existen, de hecho, trabajos experimentales que investigan el rendimiento que proporcionan en tareas de reconocimiento diferentes distancias o medidas de disimilitud definidas en el espacio vectorial de los parámetros acústicos [Nocerino, 85] [Shikano, 86].

3-. Metodología.

Para evaluar las distintas representaciones paramétricas propuestas en el capítulo anterior es necesario un criterio de calidad, que estará basado en una medida de la distorsión introducida por cada representación. Parece lógico utilizar un sistema de reconocimiento para establecer dicha medida de distorsión. Así, a partir de cada una de las representaciones paramétricas se obtiene un conjunto de modelos o patrones. Será considerada óptima aquella representación paramétrica cuyos modelos proporcionen tasas más altas de reconocimiento.

El método aplicado parte de una base de datos de palabras aisladas, con un vocabulario de tamaño N y con M muestras por palabra. Supóngase un conjunto de P parametrizaciones. Se calculan las P parametrizaciones para cada una de las N*M palabras de la base de datos.

Para incluir en el conjunto de test todas las muestras[2] de la base de datos, se aplica el método de validación cruzada [Raudys, 91], también conocido como *Leaving-K-Out* (LKOUT) [Segura, 91]. En primer lugar se consideran varias particiones de la base de datos. Cada partición consta de un conjunto de test formado por K<M muestras de cada palabra, y un conjunto de modelos formado por las muestras restantes (M-K muestras de cada palabra). Cada muestra perteneciente al conjunto de test se compara con todas las muestras del conjunto de modelos. Si la muestra modelo reconocida y la muestra de test corresponden a una misma palabra, contabilizamos un acierto. En caso contrario,

---

[2] Cada muestra o palabra parametrizada consiste en una secuencia de vectores de parámetros. Se está hablando de P parametrizaciones distintas y por tanto de P experimentos de reconocimiento.

contabilizamos un error. Para obtener las distancias entre muestras de test y muestras modelo, se utiliza una técnica denominada *dynamic time warping* (DTW) [Rabiner, 81]. La muestra modelo reconocida será aquélla cuya distancia DTW a la muestra de test sea mínima.

3.1-. *Dynamic Time Warping.*

La variabilidad temporal en la pronunciación de palabras aisladas impide una comparación directa de sus versiones parametrizadas. Previamente debe realizarse una normalización temporal de las secuencias de vectores, de forma que la comparación se realice entre secuencias de la misma longitud. El método de normalización, basado en algoritmos de programación dinámica [Bellman, 72], se denomina *Dynamic Time Warping* (DTW) [Rabiner, 81], traducido como Alineamiento Temporal No Lineal. Este método, además de la normalización temporal, proporciona una medida de disimilitud entre las palabras, lo cual permite definir un clasificador de distancia mínima.

Dadas dos palabras X=(x(1),x(2),..,x(I)) e Y=(y(1),y(2),..,y(J)), de longitudes I y J, el método de normalización establece una aplicación de alineamiento óptima F, de longitud L_F, entre las secuencias de vectores de parámetros correspondientes a dichas palabras. Dicha aplicación está dada por [Casacuberta, 92]:

$$\begin{aligned} &F: I_F \rightarrow I_X \times I_Y \\ &F(k) = \big(i(k), j(k)\big) \quad i(k) \in I_X,\ j(k) \in I_Y,\ \forall k \in I_F \end{aligned} \tag{61}$$

con $I_F=\{1,..,L_F\}$, $I_X=\{1,..,I\}$, $I_Y=\{1,..,J\}$.

Esta aplicación relaciona entre sí sucesivas parejas de vectores de las secuencias X e Y, en concreto aquellos vectores que tienen una mayor similitud según una cierta métrica o distancia local. Una vez establecida la aplicación óptima, la disimilitud entre X e Y se define como la suma normalizada de las distancias locales entre los vectores relacionados por la aplicación. La aplicación de alineamiento óptima debe cumplir dos restricciones físicas: debe ser monótona creciente y continua. También suelen aplicarse restricciones de pendiente, es decir, restricciones sobre la máxima distorsión temporal permitida.

Como métrica local d(x(i(k)),y(j(k))) suele utilizarse la distancia euclídea. El cálculo de la disimilitud entre secuencias de vectores $D_F(X,Y)$ se realiza a lo largo del camino de alineamiento. Cada elemento simple de trayectoria, es decir, el camino recorrido entre los instantes *k-1* y *k* de F, se denomina *producción*. Estrictamente, se define *producción simple positiva* como todo par (a,b) / $a,b \in Z^{\geq 0}$. Esta definición permite expresar el camino de alineamiento como una sucesión de producciones *G(k)*, las cuales deben cumplir las restricciones establecidas para conectar dos puntos en el plano IxJ [Casacuberta, 92]:

$$G(k) = \big(a(k), b(k)\big) \quad \forall k \in I_F \tag{62}$$

$$F(k) = \big(i(k-1) + a(k), i(k-1) + b(k)\big) \quad \forall k \in I_F \tag{63}$$

Cada producción introduce en el cálculo de la disimilitud un peso w(a(k),b(k)) que trata de expresar la longitud de la trayectoria elemental. La medida de disimilitud está normalizada por la longitud N del camino de alineamiento, es decir, por la suma de pesos a lo largo de dicho camino [Casacuberta, 92]:

$$D_F(X,Y) = \frac{\sum_{k=1}^{L_F} d(x(i(k)), y(j(k))) \cdot w(a(k), b(k))}{\sum_{k=1}^{L_F} w(a(k), b(k))} \qquad (64)$$

La minimización de la disimilitud exige que dicha longitud sea independiente del camino. Esto restringe las posibles definiciones de los pesos. Por ejemplo, serían definiciones válidas:

w(a,b)=a N=I
w(a,b)=b N=J
w(a,b)=a+b N=I+J

La especificación de las restricciones y los pesos de las producciones determina completamente las características del alineamiento temporal. El camino de alineamiento y la disimilitud óptimos, F_0 y D(X,Y), se obtienen minimizando la expresión (64) a lo largo de todos los caminos posibles (Fig. 15). Este proceso de búsqueda-minimización se realiza, como se ha dicho, mediante técnicas de programación dinámica [Bellman, 72], que conducen al siguiente resultado [Casacuberta, 92]:

$$\begin{aligned} &D(X,Y) = g(I,J) / N \\ &g(i,j) = \min_{\forall(a,b)} \left(g(i-a, j-b) + d(x(i), y(j)) \cdot w(a,b) \right) \\ &g(1,1) = d(x(1), y(1)) \cdot w(a(1), b(1)) \\ &w(a(1), b(1)) \quad \text{producción inicial} \end{aligned} \qquad (65)$$

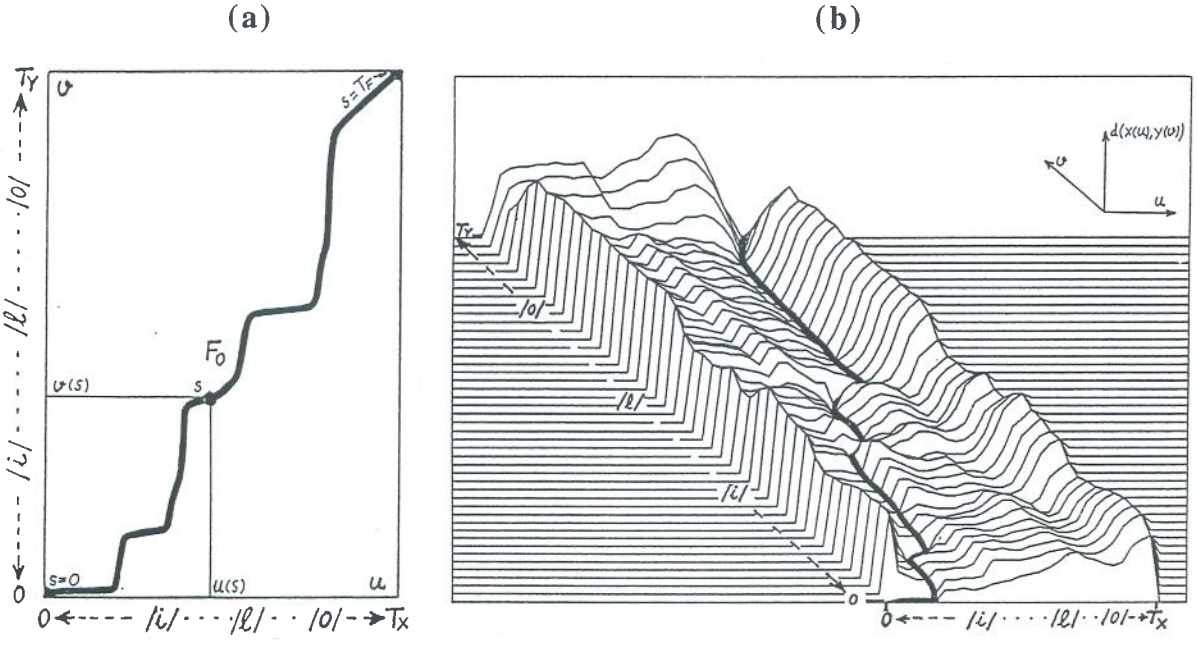

Figura 15. (a) Función de alineamiento temporal en el plano de tiempos u-v correspondiente a dos pronunciaciones de la palabra /ilo/. (b) Superficie de distancias locales entre los sucesivos vectores de parámetros de ambas pronunciaciones, con el camino de mínima distancia sobreimpresionado [Casacuberta, 92].

3.2-. Base de datos DIG2.

La base de datos de palabras aisladas utilizada en estos experimentos (DIG2) ha sido cedida por el Grupo de Reconocimiento de Formas e Inteligencia Artificial (GRFIA) del Departamento de Sistemas Informáticos y Computación (DSIC) de la Universidad

Politécnica de Valencia. DIG2 contiene pronunciaciones de los dígitos castellanos del cero al nueve. Se han utilizado 10 locutores, 5 mujeres (L0-L4) y 5 hombres (L5-L9), los cuales han realizado 10 repeticiones de cada dígito, sumando un total de 1000 pronunciaciones.

La adquisición de señales tiene las siguientes características:

1-. Frecuencia de muestreo: 16 kHz.
2-. Frecuencia de corte del filtro antialiasing: 7556 Hz.
3-. Precisión del conversor A/D: 12 bits.
4-. Expansión de muestras a formato entero corto: 16 bits.
5-. Conversión de ficheros de señal a formato ESPS (FEA_SD).

La duración de las señales es generalmente inferior a 1 segundo. En este tiempo se incluyen la propia pronunciación y márgenes de silencio de unos 100 ms antes y después de la misma.

La frecuencia de muestreo de la base de datos utilizada por el GRAH-MBAT para la decodificación acústico-fonética (DAF) de discurso continuo es 16 kHz, mientras que los resultados comparativos examinados en la bibliografía corresponden a sistemas que operan con frecuencias de muestreo inferiores (8, 10 kHz) [Davis, 80] [Paliwal, 82a] [Partalo, 89] [Hunt, 89]. Ello incrementa el interés del experimento propuesto, ya que proporciona resultados comparativos sobre una base de datos (DIG2) muestreada a 16 kHz.

4-. Realización experimental.

El experimento fue diseñado conjuntamente por el GRAH-MBAT y el GRFIA, con el objetivo de mejorar la parametrización utilizada por estos grupos -coeficientes cepstrales calculados a partir de un banco de filtros con escala Bark-, y comparar su rendimiento frente a otras técnicas alternativas, concretamente frente a parámetros derivados de un análisis de predicción lineal. Se ha comprobado la influencia de los siguientes factores: preénfasis, transformación bilineal, inclusión de la energía y longitud del vector de parámetros.

La idea experimental se basó en los trabajos de Davis y Mermelstein [Davis, 80] y de K.F. Lee [Lee, 89]. También se han tenido en cuenta los trabajos de K.K. Paliwal [Paliwal, 82a] [Paliwal, 82b] y [Paliwal, 84]. Se han utilizado máquinas Sun IPC y Sparc 2, en entorno UNIX-XWindows, y el lenguaje de programación C. Se ha hecho uso extensivo de comandos y rutinas ESPS [ESPS, 93], y se han utilizado formatos de fichero ESPS para las señales y las parametrizaciones, con objeto de generalizar el proceso de experimentación. Asimismo, en el desarrollo y monitorización de los experimentos se ha utilizado el programa gráfico XWAVES, asociado al ESPS.

4.1-. Preprocesamiento.

La etapa de preprocesamiento tiene las siguientes características:

Longitud de tramo	32 ms
Solapamiento	16 ms
Ventana	Hamming

Alternativamente, se ha aplicado preénfasis:

Coeficiente de preénfasis	0.9375

4.2-. Parámetros acústicos.

Se ha experimentado, por un lado, con parámetros obtenidos de un análisis espectral mediante transformadas de Fourier (coeficientes de un banco de filtros con escala Bark, cepstrales a partir de dicho banco de filtros y cepstrales a partir de FFT), y por otro lado, con parámetros obtenidos de un análisis de predicción lineal (coeficientes de reflexión, coeficientes LAR y cepstrales a partir de LP). Precisamente, uno de los objetivos del experimento es comprobar cuál de los dos métodos de análisis proporciona mejores resultados, y en qué condiciones (Tabla II).

Las longitudes asignadas inicialmente a los vectores de cepstrales y de parámetros LP han sido tomadas del trabajo de K.F. Lee [Lee, 89]. No se ha aplicado ventana de *liftering*. Se han generado vectores con y sin preénfasis, y en el caso de los cepstrales obtenidos a partir de LP y FFT, se ha tratado de estimar el valor óptimo del coeficiente de la transformación bilineal, experimentando con varios valores (a=0, 0.4, 0.5, 0.6, 0.7 y 0.8).

Los parámetros RC, LAR, LPCEP y FFTCEP se calculan mediante un comando ESPS. Los coeficientes BFB se calculan mediante promediado en bandas de una FFT, según el método descrito en el apartado 2.3 de este informe. A partir de ellos se extraen también los coeficientes cepstrales (BFBCEP). En todos los casos, la longitud de la FFT coincide con el tamaño de la ventana de análisis: 16 ms, es decir, 512 muestras.

Tabla II. Conjunto de parámetros puestos a prueba en el experimento de reconocimiento.

PARAMETROS	Longitud del vector	BLT
1-. Coeficientes de Reflexión (RC)	14	NO
2-. Log-Area Ratios (LAR)	14	NO
3-. Coeficientes Cepstrales a partir de LP (LPCEP)	12	SI
4-. Coeficientes Cepstrales a partir de FFT (FFTCEP)	12	SI
5-. Coeficientes de Banco de Filtros con escala Bark (BFB)	21	NO
6-. Coeficientes Cepstrales a partir de BFB (BFBCEP)	12	NO

4.3-. Estadísticas.

Examinadas las tasas de reconocimiento de estas parametrizaciones, se han escogido los tres vectores de cepstrales (LPCEP, FFTCEP y BFBCEP) para comprobar la influencia de la energía y de la longitud del vector de parámetros en dichas tasas. Se han generado vectores de longitudes 6, 8, 10 y 12, y partir de éstos, se ha obtenido un nuevo conjunto de vectores, añadiendo como primera componente el logaritmo de la energía de la señal, que se reescala para adaptar su varianza a la varianza máxima de cada vector (Tabla IV).

Se han tomado muestras de 100 ficheros de la base de datos (una repetición al azar de cada dígito por cada locutor), obteniendo la media y la varianza globales de cada parámetro en dichos ficheros (Figura 16, columna izquierda). Se han calculado también las medias y las varianzas locales de los parámetros en cada fichero, obteniendo 100 valores distintos de media y varianza por cada parámetro. Estos datos permiten estimar la variabilidad conjunta interlocutor e interdígito de las medias y las varianzas (Figura 16, columna derecha). Para obtener la razón de escalado, se han calculado también la media y la varianza del logaritmo de la energía. Finalmente se han obtenido histogramas de los parámetros acústicos y de la energía en los 100 ficheros de la muestra (Figuras 17, 18, 19 y 20).

Tabla III. Medias y desviaciones típicas de la energía y de los coeficientes cepstrales, obtenidas a partir de una muestra de 100 ficheros de la base de datos DIG2.

Parámetro	Componente	Media	Desviación típica
LOG_ENERGIA	0	9.72	3.32
LOG_BFB_ENERGIA	0	18.45	3.32
BFBCEP	0	-1.45	1.49
BFBCEP	1	0.31	1.16
BFBCEP	2	-0.1	0.66
BFBCEP	3	-0.53	1
BFBCEP	4	-0.1	0.57
BFBCEP	5	-0.12	0.45
BFBCEP	6	0.02	0.42
BFBCEP	7	0.08	0.34
BFBCEP	8	-0.02	0.32
BFBCEP	9	-0.03	0.27
BFBCEP	10	-0.1	0.31
BFBCEP	11	-0.12	0.23
FFTCEP	0	-0.08	0.36
FFTCEP	1	0.04	0.23
FFTCEP	2	0.05	0.16
FFTCEP	3	-0.08	0.12
FFTCEP	4	0.04	0.09
FFTCEP	5	-0.04	0.08
FFTCEP	6	0	0.06
FFTCEP	7	-0.01	0.06
FFTCEP	8	-0.01	0.08
FFTCEP	9	-0.02	0.08
FFTCEP	10	-0.01	0.07
FFTCEP	11	-0.02	0.07
LPCEP	0	0.12	1.07
LPCEP	1	0.09	0.42
LPCEP	2	0.17	0.33
LPCEP	3	-0.16	0.24
LPCEP	4	0.06	0.16
LPCEP	5	-0.01	0.16
LPCEP	6	-0.01	0.12
LPCEP	7	0	0.09
LPCEP	8	-0.01	0.08
LPCEP	9	0	0.08
LPCEP	10	-0.01	0.07
LPCEP	11	-0.01	0.07

Tabla IV. Razones de reescalado que resultan para el logaritmo de la energía con respecto a la máxima desviación típica de los vectores de parámetros.

máxima_dev (BFBCEP) / dev (LOG_BFB_ENERGIA):	0.4488
máxima_dev (FFTCEP) / dev (LOG_ENERGIA):	0.1084
máxima_dev (LPCEP) / dev (LOG_ENERGIA):	0.3223

Sólo los primeros coeficientes cepstrales -aproximadamente los 6 primeros- tienen una varianza apreciable y contribuyen a la función distancia en proporción mucho mayor que el resto de coeficientes. De ahí que no puedan esperarse grandes mejoras al aumentar la longitud del vector de parámetros. Por otra parte, las medias de las primeros cepstrales parecen depender del método de análisis, mientras que las de los últimos en todos los casos tienden a cero o a valores ligeramente negativos. En cuanto a la energía, insertada con varianza igual a la máxima varianza de cada vector de parámetros, su contribución a la función distancia es importante, y parece lógico que introduzca una mejora significativa en las tasas de reconocimiento. De hecho, los patrones de energía pueden utilizarse como criterio para discriminar entre un grupo reducido de palabras como los dígitos.

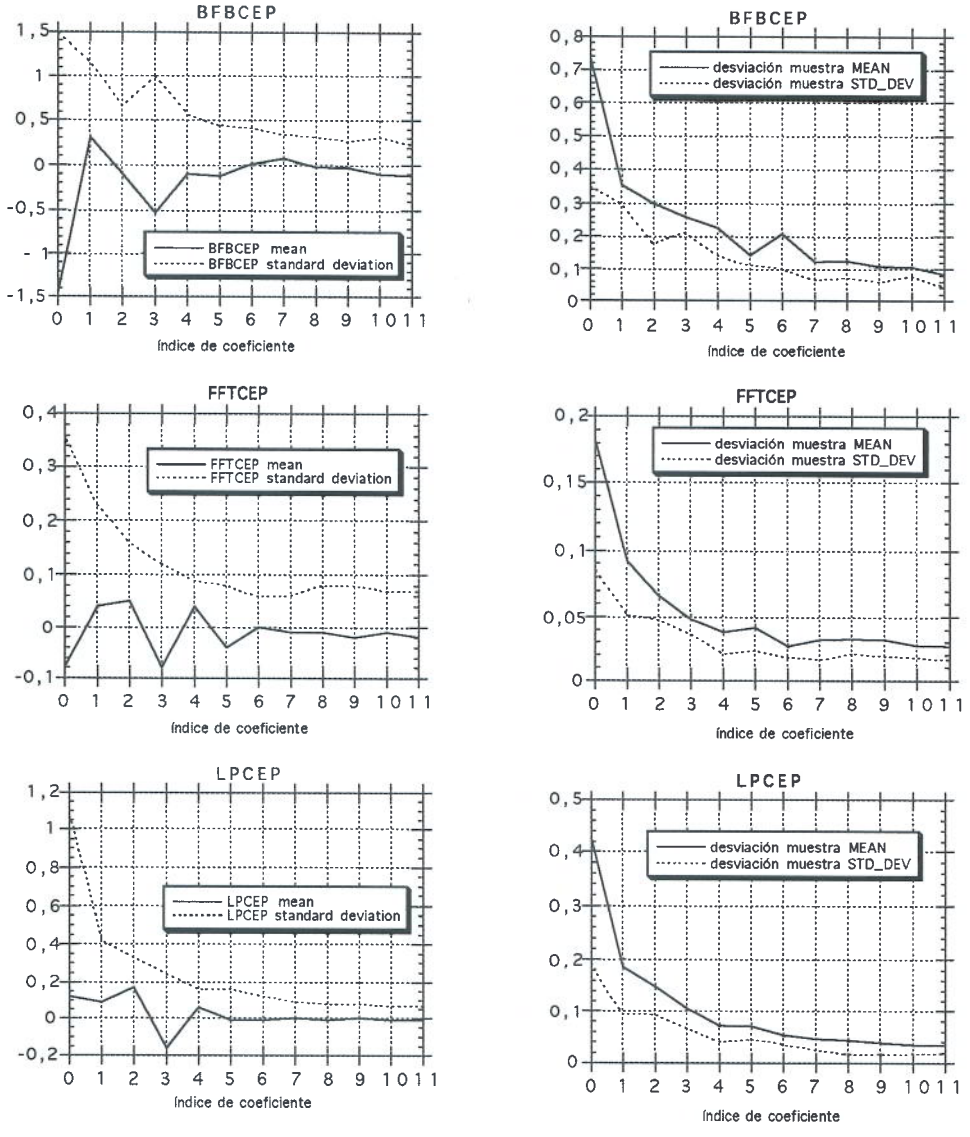

Figura 16. Media y desviación típica globales de los vectores BFBCEP, FFTCEP y LPCEP (columna izquierda), y varianzas de ambos datos en los 100 ficheros de la muestra de DIG2 utilizada en la estadística (columna derecha).

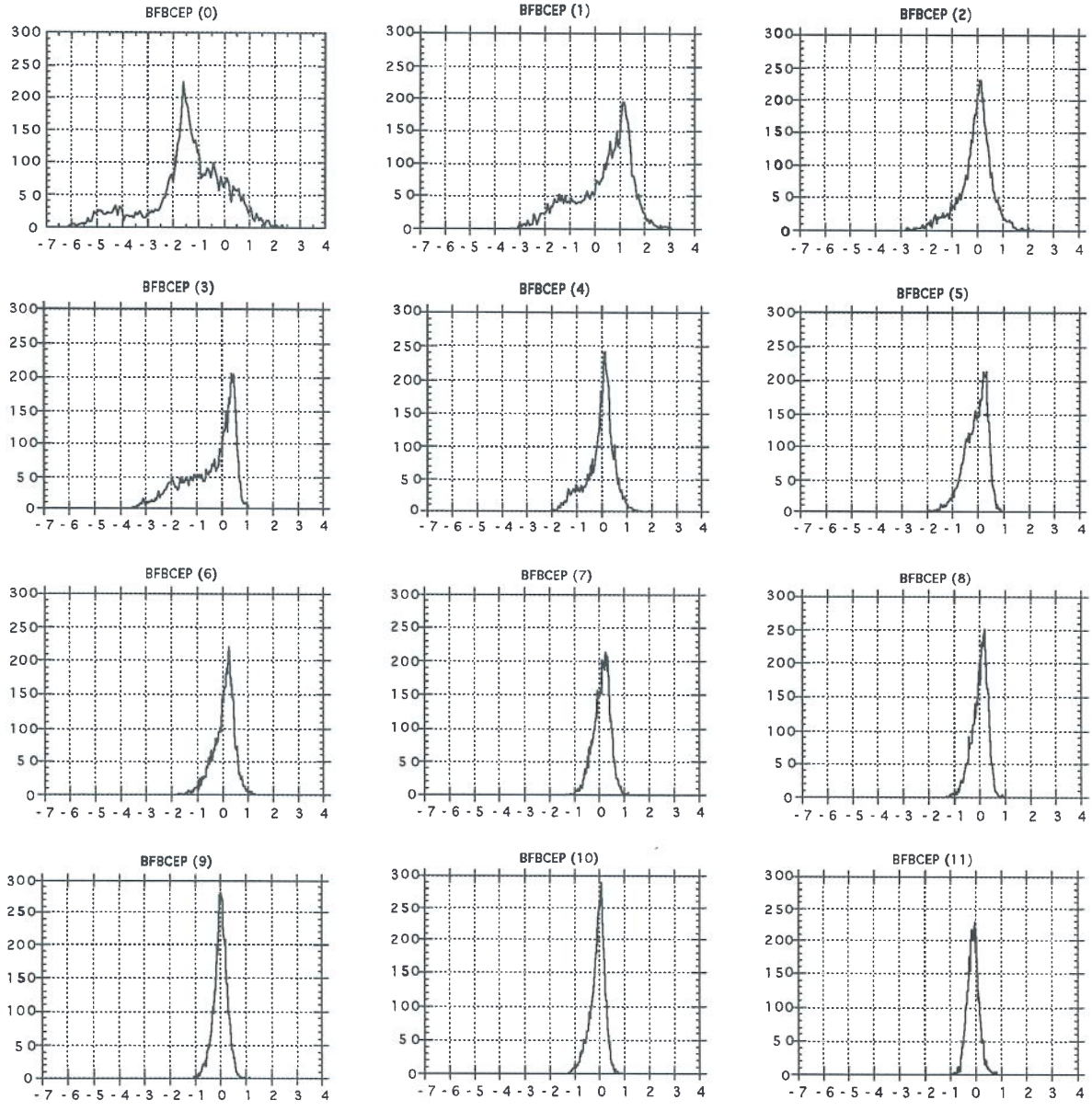

Figura 17. Histogramas de las 12 componentes del vector BFBCEP.

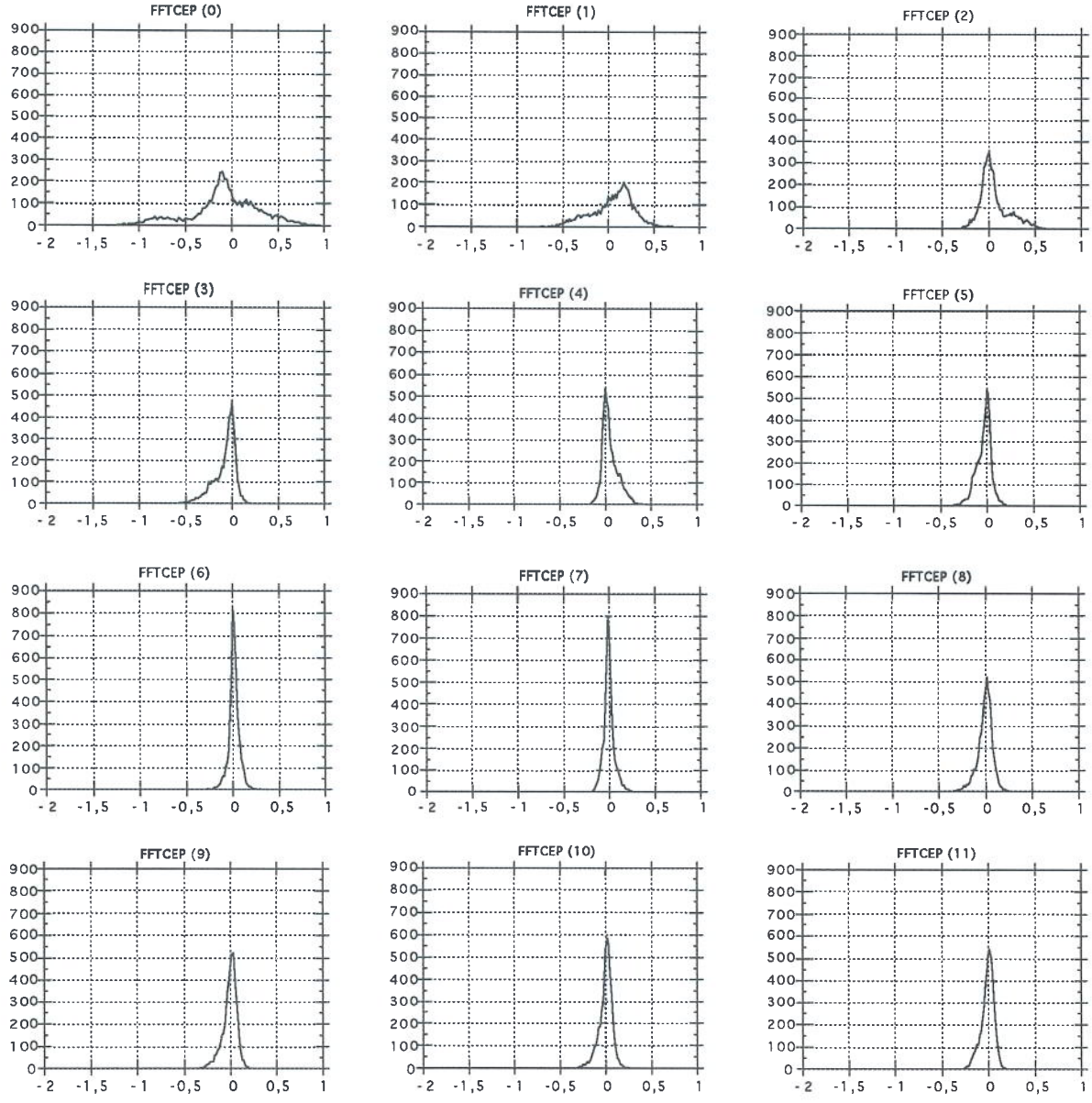

Figura 18. Histogramas de las 12 componentes del vector FFTCEP.

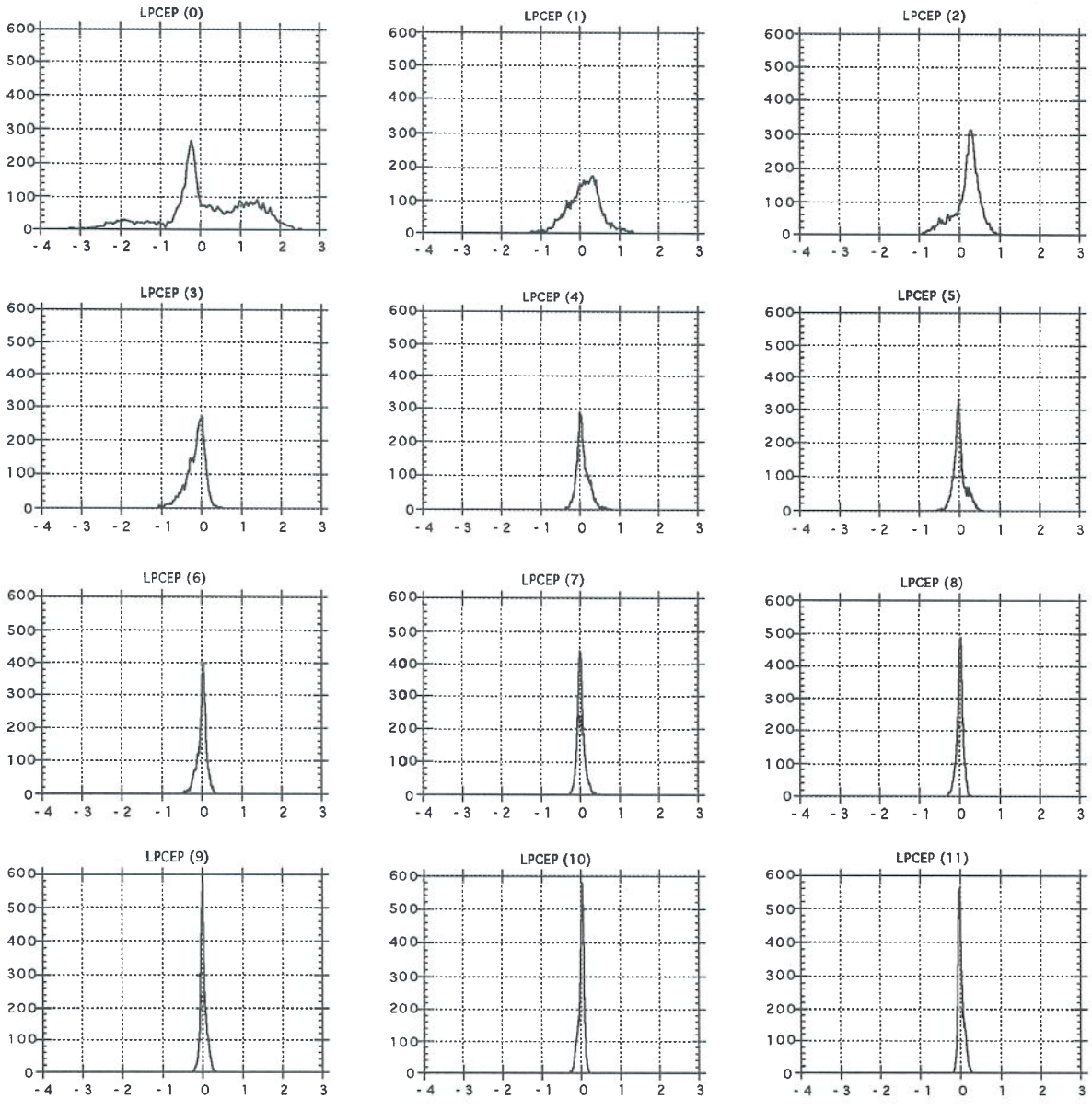

Figura 19. Histogramas de las 12 componentes del vector LPCEP.

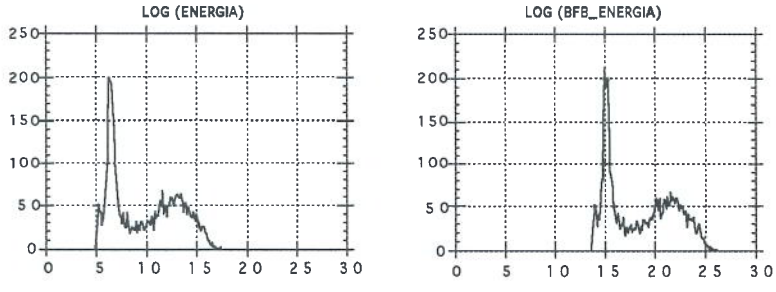

Figura 20. Histogramas de LOG(ENERGIA) y LOG(BFB_ENERGIA).

5-. Resultados.

A continuación se comentan brevemente los resultados del experimento. En primer lugar se han puesto a prueba todas las representaciones paramétricas sobre un conjunto de 5 particiones L2OUT de DIG2. A continuación se ha repetido el experimento sobre el conjunto de particiones L8OUT que resulta de invertir los conjuntos modelo y test en cada una de las particiones L2OUT anteriores. Finalmente se han escogido BFBCEP, FFTCEP y LPCEP para experimentar con la longitud del vector de parámetros y con la inclusión de la energía, utilizando las mismas particiones L2OUT y L8OUT. Cada caso queda ilustrado con una tabla y una gráfica donde se muestran las tasas de reconocimiento.

5.1-. Experimento L2OUT.

Se extrae de la base de datos un corpus de entrenamiento (CE) formado por las pronunciaciones de 8 locutores, dejando como corpus de test (CT) las pronunciaciones de los 2 restantes, correspondientes a un locutor femenino y a un locutor masculino (Tabla V). Esta operación se realiza 5 veces, para que todos los locutores aparezcan en el corpus de test. La combinación hombre-mujer de cada partición se ha escogido al azar.

Tabla V. Particiones de la base de datos DIG2 para el experimento L2OUT.

# Partición	CE	CT
1	L1 L2 L3 L4 L5 L6 L8 L9	L0 L7
2	L0 L2 L3 L4 L5 L6 L7 L8	L1 L9
3	L0 L1 L3 L4 L5 L7 L8 L9	L2 L6
4	L0 L1 L2 L4 L5 L6 L7 L9	L3 L8
5	L0 L1 L2 L3 L6 L7 L8 L9	L4 L5

Tabla VI. Tasas de reconocimiento mediante DTW para los vectores de parámetros BFB, BFBCEP, FFTCEP, LPCEP, RC y LAR. En los casos FFTCEP y LPCEP se ha experimentado con distintos valores del coeficiente de la transformación bilineal. Experimento L2OUT.

Parámetro	Preénfasis	BLT	% Rec	Parámetro	Preénfasis	BLT	% Rec
BFB	NO	_	99.2	LPCEP	NO	_	95.8
BFB	SI	_	99.2	LPCEP	NO	0.4	97.7
BFBCEP	NO	_	99.9	LPCEP	NO	0.5	98.1
BFBCEP	SI	_	99.9	LPCEP	NO	0.6	98.9
FFTCEP	NO	_	96.6	LPCEP	NO	0.7	99.1
FFTCEP	NO	0.4	98.1	LPCEP	NO	0.8	98.9
FFTCEP	NO	0.5	98.4	LPCEP	SI	_	95.9
FFTCEP	NO	0.6	98.3	LPCEP	SI	0.4	98.6
FFTCEP	NO	0.7	97.6	LPCEP	SI	0.5	98.9
FFTCEP	NO	0.8	97.9	LPCEP	SI	0.6	99.3
FFTCEP	SI	_	96.2	LPCEP	SI	0.7	99.6
FFTCEP	SI	0.4	98.1	LPCEP	SI	0.8	99.2
FFTCEP	SI	0.5	98.3	RC	NO	_	93.2
FFTCEP	SI	0.6	98.4	RC	SI	_	94.4
FFTCEP	SI	0.7	97.5	LAR	NO	_	96.3
FFTCEP	SI	0.8	97.5	LAR	SI	_	94.5

Los resultados de reconocimiento mediante DTW con distancia euclídea (Tabla VI) indican una clara superioridad de los coeficientes BFBCEP (99.9%). No obstante, los coeficientes LPCEP con BLT y preénfasis alcanzan tasas de reconocimiento similares (99.6%), por encima de BFB (99.2%). El vector FFTCEP proporciona tasas ligeramente más bajas (98.4%). Los vectores RC y LAR presentan, por su parte, tasas muy inferiores (94.4% y 96.3%, respectivamente). El preénfasis de la señal de voz, que aparentemente no afecta a las tasas de BFB y BFBCEP, y escasamente a las de FFTCEP, mejora ostensiblemente las tasas de LPCEP. En cuanto al coeficiente de la transformación bilineal, tanto FFTCEP como LPCEP muestran valores óptimos, FFTCEP en a=0.6 y LPCEP en a=0.7 (Figura 21).

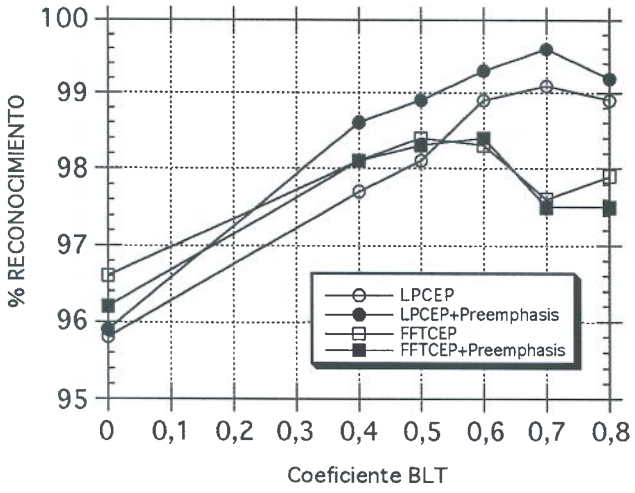

Figura 21. Tasas de reconocimiento mediante DTW para los vectores de parámetros FFTCEP y LPCEP en función del coeficiente de la transformación bilineal. Experimento L2OUT.

5.2-. Experimento L8OUT.

La escasa dificultad de la tarea y la eficacia del algoritmo de comparación proporcionan tasas demasiado altas y poco discriminativas. Para aumentar las diferencias entre las tasas de reconocimiento, se ha optado por dificultar artificialmente la tarea disminuyendo el número de modelos por dígito en el algoritmo DTW.

En cada partición se extrae de la base de datos un corpus de entrenamiento (CE) formado por las pronunciaciones de 2 locutores, correspondientes a un locutor femenino y a un locutor masculino, dejando como corpus de test (CT) las pronunciaciones de los 8 locutores restantes (Tabla VII). Se realizan 5 particiones, invirtiendo los conjuntos modelo y test del experimento L2OUT. De esta forma todos los locutores aparecen alternativamente en el corpus de entrenamiento. Se tienen en total 4000 muestras de test.

Como resultado de estos cambios, las tasas de reconocimiento disminuyen (Tabla VIII). El vector BFBCEP mantiene la tasa más alta (98.65%), seguido de LPCEP (97.97%) y de BFB (97.82%), mientras que FFTCEP se encuentra muy por debajo (93.95%). Se constata que aplicar preénfasis sólo mejora el rendimiento de los parámetros derivados de análisis LP. Por otra parte, el vector LPCEP no presenta un valor óptimo del coeficiente BLT dentro del rango de valores examinado -sus tasas aumentan monótonamente y dan un máximo en a=0.8. El vector FFTCEP, en cambio, presenta un valor óptimo en a=0.7 (Figura 22). En experimentos posteriores se utilizarán como valores óptimos del coeficiente BLT los obtenidos en el experimento L2OUT.

Tabla VII. Particiones de la base de datos DIG2 para el experimento L8OUT.

# Partición	CE	CT
1	L0 L7	L1 L2 L3 L4 L5 L6 L8 L9
2	L1 L9	L0 L2 L3 L4 L5 L6 L7 L8
3	L2 L6	L0 L1 L3 L4 L5 L7 L8 L9
4	L3 L8	L0 L1 L2 L4 L5 L6 L7 L9
5	L4 L5	L0 L1 L2 L3 L6 L7 L8 L9

Tabla VIII. Tasas de reconocimiento mediante DTW para los vectores de parámetros BFB, BFBCEP, FFTCEP y LPCEP. En los casos FFTCEP y LPCEP se ha experimentado con distintos valores del coeficiente de la transformación bilineal. Experimento L8OUT.

Parámetro	Preénfasis	BLT	% Rec	Parámetro	Preénfasis	BLT	% Rec
BFB	NO	_	97.82	LPCEP	NO	_	89.85
BFB	SI	_	97.75	LPCEP	NO	0.4	95.38
BFBCEP	NO	_	98.65	LPCEP	NO	0.5	96.35
BFBCEP	SI	_	98.47	LPCEP	NO	0.6	96.97
FFTCEP	NO	_	90.00	LPCEP	NO	0.7	97.28
FFTCEP	NO	0.4	93.03	LPCEP	NO	0.8	97.57
FFTCEP	NO	0.5	93.82	LPCEP	SI	_	90.40
FFTCEP	NO	0.6	93.88	LPCEP	SI	0.4	95.72
FFTCEP	NO	0.7	93.95	LPCEP	SI	0.5	96.65
FFTCEP	NO	0.8	92.80	LPCEP	SI	0.6	97.28
FFTCEP	SI	_	89.50	LPCEP	SI	0.7	97.75
FFTCEP	SI	0.4	92.82	LPCEP	SI	0.8	97.97
FFTCEP	SI	0.5	93.57				
FFTCEP	SI	0.6	93.80				
FFTCEP	SI	0.7	93.93				
FFTCEP	SI	0.8	92.47				

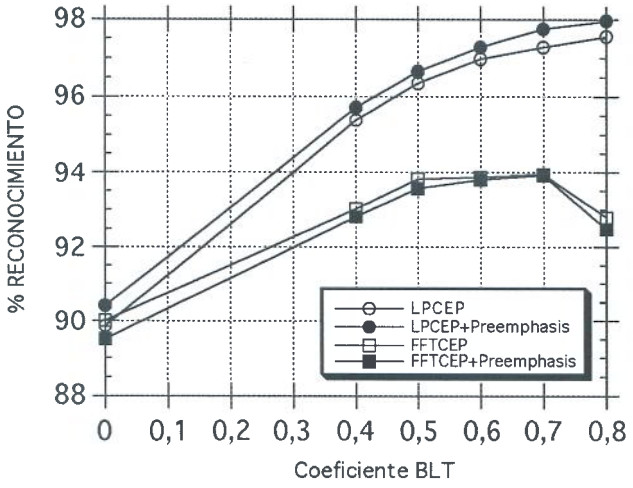

Figura 22. Tasas de reconocimiento mediante DTW para los vectores de parámetros FFTCEP y LPCEP en función del coeficiente de la transformación bilineal. Experimento L8OUT.

5.3-. Tamaño del vector de parámetros y energía.

Finalmente, interesa comprobar cómo evolucionan las tasas de reconocimiento en función del tamaño (L) del vector de parámetros. Reducir el tamaño del vector significa reducir, aún en mayor medida, el coste computacional en las etapas de cuantificación vectorial y estimación de modelos acústicos. Para ello, se ha planteado un nuevo experimento, sobre la misma base de datos, utilizando únicamente los coeficientes cepstrales: BFBCEP (sin preénfasis), LPCEP (con preénfasis, a=0.7) y FFTCEP (sin preénfasis, a=0.6). Alternativamente se ha añadido, como componente adicional, el logaritmo de la energía, ya que también interesa conocer su influencia en el reconocimiento, y se ha reescalado de forma que su varianza coincida con la máxima varianza del vector donde se inserta, evitando así que su inclusión distorsione la distancia euclídea definida entre los vectores de parámetros. Se han generado, por tanto, vectores de tamaños 6, 8, 10 y 12, con y sin energía.

Nuevamente se han obtenido resultados con un número elevado y con un número reducido de modelos por dígito, manteniendo las particiones L2OUT y L8OUT definidas en los apartados anteriores (Tablas IX y X, respectivamente). En todos los casos,

lógicamente, la inclusión de la energía mejora las tasas de reconocimiento. La combinación LPCEP+LOG_ENERGIA, cuyo rendimiento apenas depende del tamaño del vector, mantiene tasas elevadas en todos los casos e incluso muestra un tamaño de vector óptimo, entre L=8 y L=10. BFBCEP, por su parte, presenta tasas que aumentan monótonamente con el tamaño del vector. Finalmente, el rendimiento de FFTCEP es inferior a los anteriores en todos los casos (Figuras 23 y 24).

Tabla IX. Tasas de reconocimiento mediante DTW para los vectores de parámetros BFBCEP, FFTCEP y LPCEP. Experimento L2OUT. Se ha experimentado con diferentes longitudes del vector de parámetros, y alternativamente se ha añadido, como primera componente, el logaritmo de la energía.

Parámetro	Energía	L	% Rec	Parámetro	Energía	L	% Rec
BFBCEP	NO	6	98.9	FFTCEP	SI	6	98.7
BFBCEP	NO	8	99.2	FFTCEP	SI	8	99.1
BFBCEP	NO	10	99.6	FFTCEP	SI	10	99.3
BFBCEP	NO	12	99.9	FFTCEP	SI	12	99.7
BFBCEP	SI	6	99.5	LPCEP	NO	6	98.7
BFBCEP	SI	8	99.5	LPCEP	NO	8	99.4
BFBCEP	SI	10	99.8	LPCEP	NO	10	99.5
BFBCEP	SI	12	99.9	LPCEP	NO	12	99.6
FFTCEP	NO	6	96.4	LPCEP	SI	6	99.7
FFTCEP	NO	8	96.8	LPCEP	SI	8	99.8
FFTCEP	NO	10	98.0	LPCEP	SI	10	99.8
FFTCEP	NO	12	98.4	LPCEP	SI	12	99.7

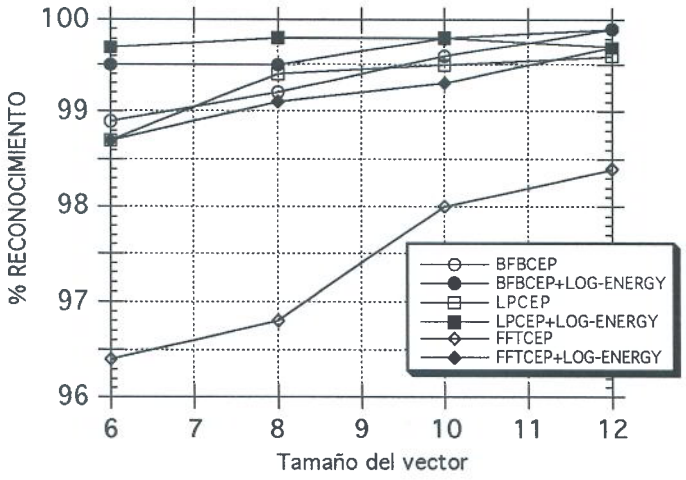

Figura 23. Tasas de reconocimiento mediante DTW para los vectores de parámetros BFBCEP, FFTCEP y LPCEP. Se ha experimentado con diferentes tamaños del vector de parámetros y se ha incluido alternativamente el logaritmo de la energía. Experimento L2OUT.

Tabla X. Tasas de reconocimiento mediante DTW para los vectores de parámetros BFBCEP, FFTCEP y LPCEP. Experimento L8OUT. Se ha experimentado con diferentes longitudes del vector de parámetros, y alternativamente se ha añadido, como primera componente, el logaritmo de la energía.

Parámetro	Energía	L	% Rec	Parámetro	Energía	L	% Rec
BFBCEP	NO	6	97.38	FFTCEP	SI	6	96.85
BFBCEP	NO	8	97.68	FFTCEP	SI	8	96.30
BFBCEP	NO	10	98.10	FFTCEP	SI	10	96.97
BFBCEP	NO	12	98.47	FFTCEP	SI	12	96.90
BFBCEP	SI	6	98.05	LPCEP	NO	6	97.55
BFBCEP	SI	8	98.22	LPCEP	NO	8	97.85
BFBCEP	SI	10	98.30	LPCEP	NO	10	97.65
BFBCEP	SI	12	98.62	LPCEP	NO	12	97.75
FFTCEP	NO	6	93.68	LPCEP	SI	6	98.55
FFTCEP	NO	8	93.65	LPCEP	SI	8	98.62
FFTCEP	NO	10	94.05	LPCEP	SI	10	98.62
FFTCEP	NO	12	93.80	LPCEP	SI	12	98.62

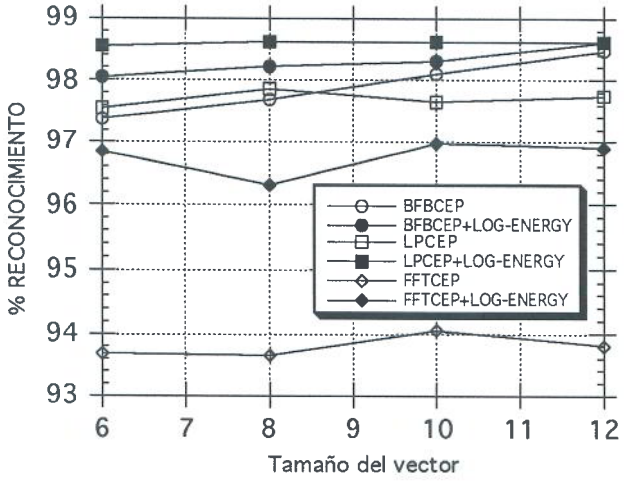

Figura 24. Tasas de reconocimiento mediante DTW para los vectores de parámetros BFBCEP, FFTCEP y LPCEP. Se ha experimentado con diferentes tamaños del vector de parámetros y se ha incluido alternativamente el logaritmo de la energía. Experimento L8OUT.

6-. Conclusiones.

La literatura presenta resultados concluyentes en cuanto a la utilización de los coeficientes cepstrales como parámetros acústicos [Davis, 80] [Paliwal, 82a] [Paliwal, 82b]. Existen dos formulaciones principales. La primera de ellas consiste en analizar la señal de voz mediante los coeficientes de un banco de filtros pasabanda cuyas anchuras están escaladas para reproducir la resolución en frecuencia, no lineal, del oído humano (escala Bark de bandas críticas: *Bark-scaled Filter Bank*, BFB). Habitualmente, se realiza mediante submuestreo-promediado en bandas críticas de una Transformada Rápida de Fourier (*Fast Fourier Transform*, FFT). La segunda formulación, basada en un análisis de predicción lineal (*Linear Prediction*, LP), obtiene un conjunto de coeficientes que equivale aproximadamente al espectro suavizado de la señal, e incluye normalmente una transformación bilineal del eje de frecuencias, que trata de reproducir la escala Bark. En ambos casos, obviamente, los coeficientes cepstrales constituyen una representación de la envolvente del espectro logarítmico.

Normalmente, dicha representación se trunca mediante una ventana de longitud L, denominada ventana de *liftering*, que puede ser rectangular -como en el presente trabajo-, aunque habitualmente se aplican ventanas tipo rampa o tipo seno remontado [Tohkura, 87] [Segura, 91], para robustecer el comportamiento del vector de parámetros frente a variaciones de locutor o de las condiciones del canal de transmisión. Por otra parte, la longitud de la ventana de *liftering* determina la dimensión del espacio vectorial de representación, y por tanto, el coste computacional de las etapas de procesamiento posteriores.

Nuestra preocupación se ha centrado en determinar cuál de estas representaciones (BFB o LP), y en qué condiciones, proporciona tasas de reconocimiento más elevadas, comprobando cómo afectan a dichas tasas la longitud del vector de parámetros (L), la escala Bark, la transformación bilineal, y el preénfasis de la señal de voz.

Las conclusiones del trabajo pueden resumirse en los siguientes puntos:

1-. La representación paramétrica que da tasas más altas de reconocimiento con DTW es, en general, BFBCEP. En segundo lugar se encuentra LPCEP, con tasas muy similares. Los parámetros restantes proporcionan tasas claramente inferiores.

2-. Se ha podido comprobar que la aplicación de la escala Bark mejora notablemente los resultados de reconocimiento obtenidos con parámetros derivados de un banco de filtros.

3-. La aplicación de una transformación bilineal a los coeficientes cepstrales obtenidos de una FFT o de análisis LP proporciona una distorsión de la escala espectral similar a la producida por el banco de filtros con escala Bark, y como consecuencia, ajustando el valor del coeficiente de la transformada, se consiguen tasas de reconocimiento más altas.

4-. Los parámetros BFB, BFBCEP y FFTCEP apenas se ven afectados por el preénfasis de la señal de voz. En todo caso, tal como muestran los datos de la Tabla VIII, parece más indicado no aplicarlo, ya que, de hacerlo, las tasas de reconocimiento en general disminuyen. Sólo los coeficientes cepstrales obtenidos mediante análisis de predicción lineal (LPCEP) presentan un aumento de las tasas de reconocimiento al aplicar preénfasis.

5-. La longitud del vector de parámetros afecta en diferente medida a cada parámetro. Sólo BFBCEP muestra en todos los casos un crecimiento monótono de las tasas a medida que aumenta el tamaño del vector. LPCEP parece mostrar un valor óptimo entre L=8 y L=10. Los resultados con FFTCEP difieren de un experimento a otro. Así, en el experimento L2OUT muestra un crecimiento monótono muy acusado, mientras que en el experimento L8OUT presenta oscilaciones.

6-. Finalmente, la inclusión de la energía como primera componente del vector de parámetros, con varianza igual a la máxima varianza de éste, incrementa significativamente las tasas de reconocimiento.

Para una implementación definitiva parece adecuado escoger entre BFBCEP y LPCEP. Esta elección dependerá del tamaño del vector de parámetros. LPCEP constituye la mejor opción con L=6 y L=8, y BFBCEP la mejor opción con L=10 y L=12. En cualquier caso, pueden realizarse nuevos experimentos con la ventana de liftering y con la longitud del vector [Tohkura, 87] [Segura, 91], extendiendo las representaciones mediante sucesivas derivadas o mediante coeficientes de regresión de los parámetros acústicos [Furui, 86].

7-. Referencias bibliográficas.

[Bellegarda, 90] J. R. Bellegarda and D. Nahamoo. "Tied Mixture Continuous Parameter Modelling for Speech Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, No. 12, pp. 2033-2045, 1990*

[Bellman, 72] R. Bellman. "Dynamic Programming". *Princeton University Press. 1972.*

[Casacuberta, 92] F. Casacuberta, E. Vidal. "Reconocimiento Automático del Habla". *Apuntes del curso. Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia. 1992.*

[Davis, 80] S.B. Davis, P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Trans. on ASSP, vol. 28, n. 4, pp. 357-366. August 1980.*

[ESPS, 93] "Entropic Signal Processing System Reference Manual". Version 5.0. *Entropic Research Laboratory. Washington DC. August 1993.*

[Furui, 86] S. Furui. "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum". *IEEE Trans. on ASSP, Vol. 34, n. 1, pp. 52-59. February 1986.*

[Huang, 93] X. D. Huang, H. W. Hon, M. Y. Hwang and K. F. Lee. "A comparative study of discrete, semicontinuous and continuous hidden Markov models". *Computer Speech and Language Vol. 7, pp. 359-368, 1993.*

[Hunt, 89] M.J. Hunt, C. Lefèbvre. "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech". *Proc. ICASSP-89, pp. 262-265. Glasgow. May 1989.*

[Lee, 89] K.F. Lee. "Automatic Speech Recognition. The development of the SPHINX System". *Kluwer Academic Publishers. 1989.*

[Makhoul, 75] J. Makhoul. "Linear Prediction: A tutorial review". Proc. of the IEEE, Vol. 63, n. 4, pp. 561-580. April 1975.

[Markel, 76] J.D. Markel, A.H. Gray. "Linear Prediction of Speech". *Springer-Verlag. 1976.*

[Nocerino, 85] N. Nocerino, F.K. Soong, L.R. Rabiner, D.H. Klatt. "Comparative study of several distortion measures for speech recognition". *Speech Communication, vol. 4, n. 4, pp. 317-331. December 1985.*

[Oppenheim, 72] A.V. Oppenheim, D.H. Johnson. "Discrete representation of signals". *Proc. of the IEEE, vol. 60, n. 6, pp. 681-691. June 1972.*

[Oppenheim, 89] A.V. Oppenheim, R.W. Schafer. "Discrete-Time Signal Processing". *Prentice-Hall. 1989.*

[Paliwal, 82a] K.K. Paliwal, P.V.S. Rao. "Evaluation of various linear prediction parametric representations in vowel recognition". *Signal Processing, vol. 4, n. 4, pp. 323-327. July 1982.*

[Paliwal, 82b] K.K. Paliwal. "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition". *Speech Communication. Vol. 1, n° 2, pp. 151-154. Agosto 1982.*

[Paliwal, 84] K.K. Paliwal. "Effect of preemphasis in vowel recognition performance". *Speech Communication. Vol. 3, n° 1, pp. 101-106. Abril 1984.*

[Papamichalis, 87] P.E. Papamichalis. "Practical Approaches to Speech Coding". *Prentice-Hall. 1987.*

[Partalo, 89] M. Partalo, Z. Sijercic. "Comparison of several speech signal feature parameters for automatic speech recognition". *Speech Communication, Vol. 8, n. 4, pp. 347-353. December 1989.*

[Rabiner,78] L.R. Rabiner, R.W. Schafer. "Digital Processing of Speech Signals". *Prentice-Hall. 1978.*

[Rabiner, 81] L.R. Rabiner, S.E. Levinson. "Isolated and Connected Word Recognition - Theory and Selected Applications". *IEEE Trans. on Communications, Vol. 29, n. 5, pp. 621-659. May 1981.*

[Raudys, 91] S. J. Raudys and A. K. Jain, "Small Sample Effects in Statistical Pattern Recognition: Recommendations for Practitioners and Open Problems," *IEEE Trans. on PAMI, vol. 13, n3, pp. 252-263, 1991.*

[Gómez, 87] P. Gómez, V. Rodellar. "Fundamentos computacionales del análisis y síntesis de voz". *Apuntes de un seminario. Grupo de Trabajo en Sistemas Parcor. Facultad de Informática. Universidad Politécnica de Madrid. Junio de 1987.*

[Segura, 91] J.C. Segura. "Modelos de Markov con cuantificación dependiente para reconocimiento de voz". *Tesis Doctoral. Departamento de Electrónica y Tecnología de Computadores. Universidad de Granada. Noviembre 1991.*

[Shikano, 86] K. Shikano. "Evaluation of LPC spectral matching measures for phonetic unit recognition". *Technical Report CMU-CS-86-108. Computer Science Department. Carnegie-Mellon University. February 1986.*

[Tohkura, 87] Y. Tohkura. "A weighted cepstral distance measure for speech recognition". *IEEE Trans. on ASSP, vol. 35, n. 10, pp. 1414-1422. October 1987.*

[Zwicker, 81] E. Zwicker, R. Feldtkeller. "Psychoacoustique. L'oreille, recépteur d'information". *Masson Ed. París, 1981.*