APROXIMACIÓN EXPERIMENTAL AL MEJOR COMPROMISO R-D PARA LA CODIFICACIÓN DE LA VOZ A MUY BAJO BIT-RATE

Juan Manuel López Soler

TESIS DOCTORAL

Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada

Granada, Febrero 1995

APROXIMACIÓN EXPERIMENTAL AL MEJOR COMPROMISO R-D PARA LA CODIFICACIÓN DE LA VOZ A MUY BAJO BIT-RATE

Juan Manuel López Soler

TESIS DOCTORAL

Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada

Granada, Febrero 1995

D. Antonio J. Rubio Ayuso
Profesor Titular de Electrónica del
Dpto. de Electrónica y
Tecnología de Computadores.
Universidad de Granada

CERTIFICA:

Que la presente memoria titulada "Aproximación Experimental al Mejor Compromiso R-D para la Codificación de la Voz a Muy Bajo Bit-Rate" ha sido realizada por **D. Juan Manuel López Soler** bajo mi dirección en el Departamento de Electrónica y Tecnología de Computadores de la Universidad de Granada. Esta memoria constituye la Tesis que D. Juan Manuel López Soler presenta para optar al grado de Doctor en Ciencias Físicas.

Granada, a 15 de Febrero de 1995

Fdo. Juan Manuel López Soler

Fdo.: Dr. D. Antonio José Rubio Ayuso
Director de la Tesis

AGRADECIMIENTOS

Esta página quiero especialmente dedicarla a agadecer a todos los que de alguna manera han colaborado en la elaboración de este trabajo, sin los que ciertamente su realización no hubiera sido posible.

A *Antonio Rubio*, director de esta tesis, que desde los comienzos supo despertar la curiosidad y motivación suficientes para justificar años de trabajo. Gracias por tus imprescindibles comentarios, sugerencias e innumerables corecciones. Gracias por haberme contagiado el placer y vertigo que proporciona la investigación.

A *Jose Carlos, M. Carmen, Jose Luis, Antonio, Victoria, Pedro, Jesús y Angel*, amigos y miembros del grupo de investigación GiPDSyC que de una forma directa pueden considerarse extrictamente co-autores de este trabajo. Gracias por la ayuda que desinteresadamente me habeis prestado, a los ya doctores por el ejemplo dado y ánimo para los que en breve tendreis que pasar por este mismo trance.

Al profesor *Nariman Farvardin* de la Universidad de Maryland, que durante dos ocasiones me permitió conocer de cerca su grupo de investigación, y por supuesto, gracias por el valioso tiempo que me dedicó en aquellas constructivas discusiones, esenciales para este trabajo. Gracias, también a sus estudiantes.

Gracias a todos los miembros del Departamento de Electrónica y Tecnología de Computadores.

Y a todos los demás amigos, gracias, especialmente a mis padres y hermanos.

Y gracias a *Nani* que sin hablar nunca de cómo reducir la redundancia, ha contribuido decisivamente a ello.

A Nani, M. Carmen y Salvador

Índice General

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

En este capítulo introductorio se formula el problema que ha motivado toda la investigación realizada; se citan algunas posibles aplicaciones de los procedimientos desarrollados, a la vez que se comentan las bases teóricas, así como una posible clasificación de los codificadores, para con ello situar el trabajo realizado. Se resumen cualitativamente los esquemas que constituyen la principal aportación y por último se esboza por capítulos el contenido del presente trabajo de investigación.

1.1 Introducción y Formulación del Problema. Aplicaciones

La necesidad de desarrollar codificadores de voz a muy baja velocidad de transmisión (*bit-rate*) para optimizar con ello el uso del canal, así como el desafío de aproximar experimentalmente el límite teórico de la curva velocidad de transmisión frente a distorsión, especialmente en la región de muy bajos *bit-rates*, han constituido la motivación principal del trabajo realizado que aquí se resume.

Para codificar óptimamente es necesario eliminar toda la redundancia existente en la señal generada por la fuente, pero no sólo eso, ya que un codificador óptimo debe conseguir la representación digital de la señal de entrada que implique el menor *bit-rate* junto con la mínima pérdida posible de *calidad.* Además de éstos, otros factores determinantes en el diseño son la *complejidad* del sistema y el *retardo* introducido. Por lo tanto, estrictamente hablando, el diseño óptimo consiste en la resolución de un problema tetradimensional que proporcione el mejor compromiso entre las cuatro magnitudes aquí citadas.

Mientras que en situaciones prácticas, cuestiones tales como la complejidad, y *bit-rate* fijo son importantes, sería deseable conocer cuál es el compromiso límite alcanzable si tales restricciones no estuvieran presentes. Éste es esencialmente el problema estudiado en este trabajo, donde se han considerado retardos máximos inferiores a 200 ms, justificados por la definición del nuevo estándar a 2400 bps en [Welch93]. En particular, se demostrará experimentalmente que combinando adecuadamente los procesos de cuantización e interpolación se puede aproximar dicho compromiso.

A pesar de la aparición de medios de transmisión ópticos con ancho de banda casi ilimitado (varios ordenes de magnitud por encima del ancho de banda de Nyquist para la señal de voz a la frecuencias típicas de muestreo) y de dispositivos de memoria con una relación coste-capacidad-tiempo de acceso cada vez mejor, la reducción del bit rate sigue siendo una cuestión clave para la transmisión y el almacenamiento

digital de la señal de voz, debido a la necesidad creciente de usar medios de ancho de banda limitado (tales como canales de radio y enlaces por satélite), a la vez de dispositivos integrados (chips) de memoria. Aplicaciones típicas de las contribuciones aquí aportadas serán de utilidad para las comunicaciones personales móviles, *buscas* o "paging", contestadores automáticos digitales, agendas electrónicas con voz, etc.

En las últimas décadas, se ha propuesto un considerable número de sistemas para llevar a cabo la tarea de la codificación de la voz a bajo *bit-rate*; casi invariablemente, se ha adoptado la predicción lineal (LPC) [Mark76] para modelar el proceso de generación de la señal de voz. Suponiendo que los procesos de excitación y filtrado en el aparato vocal sean estadísticamente independientes, dichos procesos pueden ser tratados independientemente a la hora de ser codificados. Bajo esta premisa, el trabajo que se propone se centrará en codificar la caracterización espectral del filtro vocal y posteriormente codificar la excitación de dicho filtro. Para codificar ambos procesos se porpondrá una aproximación experimental; proporcionándose un procedimiento de *bit-rate* variable que, como se evidenciará tras las evaluaciones realizadas, obtiene el mejor compromiso de entre los sistemas o procedimientos simulados, en la región de muy bajo *bit-rate*.

La utilidad e interés de la aproximación y restricciones adoptadas no sólo reside en sí misma, por las posibles aplicaciones sugeridas, sino que investigaciones desarrolladas en esta dirección pueden proporcionar un conocimiento mejor de las propiedades estructurales de la señal de voz, así como el desarrollo de modelos que pueden desembocar en mejores sistemas de reconocimiento y codificación de la voz. Problemas estos aun abiertos para la comunidad científica.

Para finalizar con esta introducción una cuestión terminológica. A lo largo del trabajo se usan en muchas ocasiones indistintamente los términos *cuantizar* y *codificar*. Aunque evidentemente son conceptos diferentes, influidos por la ambigüedad con la que ambos términos son a veces utilizados en la bibliografía, aquí se usarán de igual manera, aludiendo al proceso de representar (y a veces transmitir) digitalmente la información contenida en la señal de entrada.

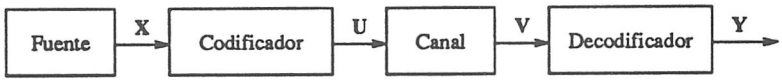

Figura 1.1: Modelo de un Sistema de Comunicación.

1.2 Antecedentes

1.2.1 La Teoría *Rate-Distorsión*

Las bases teóricas para el desarrollo de los codificadores de la fuente fueron formuladas inicialmente por Shannon, dando lugar a un campo de investigación conocido por *teoría rate-distorsión* [Berger71], [Viter79] o *teoría de la codificación de la fuente* [Gray90].

Si la fuente es ergódica y estacionaria, existe una función monótonamente no creciente, notada por $D(R)$, la cuál proporciona un límite inferior D para la distorsión promedio introducida, dada una velocidad de transmisión de R bits por muestra. Igualmente es posible considerar la función inversa de la anterior, notada por $R(D)$, definida como el valor mínimo de la información mutua promedio

$$R(D) = I(\mathbf{X};\mathbf{Y}) = \lim_{N\to\infty} I_N(\mathbf{X};\mathbf{Y}) \tag{1.1}$$

donde $\mathbf{X}$ y $\mathbf{Y}$ representan a los vectores de dimensión N de entrada y salida (ver figura (1.1)) que implican una distorsión D.

La información mutua promedio se define como el valor esperado de la cantidad de información transferida entre los dos procesos discretos en cuestión, (tomando el bit como unidad)

$$I(\mathbf{X};\mathbf{Y}) = \sum_{i=1}^{M}\sum_{j=1}^{M} P(x_i, y_j)\log_2 \frac{P(y_j|x_i)}{P(y_j)} \tag{1.2}$$

donde $P(x_i, y_j)$ es la probabilidad conjunta de ambos sucesos y $P(y_j|x_i)$ es la probabilidad condicional. La información mutua se puede interpretar como la incertidumbre a la salida de la fuente menos la incertidumbre en la salida de la fuente

conocida la salida del decodificador, ya que

$$I_N(\mathbf{X};\mathbf{Y}) = H_N(\mathbf{X}) - H_N(\mathbf{X}|\mathbf{Y}) \tag{1.3}$$

donde $H_N(\mathbf{X})$ es la entropía de orden N de la fuente y $H_N(\mathbf{X}|\mathbf{Y})$ es la entropía condicional de orden N habiendo observado Y. La función $R(D)$, definida anteriormente (expresión (1.1)), es conocida como *función rate-distorsión.*

Una de las aportaciones definitivas del desarrollo de este cuerpo teórico es el teorema de la codificación de la fuente, que se puede enunciar de la siguiente manera:

> *existe una asignación entre los símbolos del alfabeto de entrada* ($\{X\}$) *a las palabras del código* ($\{U\}$), *tal que para una distorsión promedio* D *dada, son suficientes* $R(D)$ *bits por muestra para reconstruir las muestras de la fuente en el decodificador con una distorsión promedio arbitrariamente próxima a* D.

Aunque la función $R(D)$ proporciona la velocidad teórica efectiva con la que la fuente genera información sujeta a la constricción de tolerar como máximo una distorsión promedio D, resulta ser, que en casos prácticos, dicha función se utiliza casi exclusivamente como una referencia teórica, sin más implicaciones que la de ser un límite a aproximar. Las razones de esta última afirmación residen en que

- La teoría rate-distorsión es *no constructiva*, es decir, se proporciona el límite que cabría esperar en el diseño de un sistema de codificación, pero no se proporciona el procedimiento para alcanzarlo.
- Se necesita un modelo estadístico para la fuente que no siempre es lo suficientemente válido (es conocida la dificultad de caracterizar la señal de voz con una función distribución de probabilidad que sea tratable matemáticamente).
- No se tiene en cuenta que el receptor final es un humano que juzgará la distorsión promedio con criterios subjetivos, difícilmente incorporables en el desarrollo teórico.

Toda la argumentación anterior justifica que para una fuente dada, aun pudiéndose conocer el límite teórico rate-distorsión, es conveniente proporcionar procedimientos para aproximar experimentalmente dicho límite. Ésta fué una de las motivaciones que dieron lugar al trabajo de investigación que aquí se resume.

1.2.2 Clasificación de los Codificadores

Para concluir con los antecedentes que han motivado este trabajo, en este apartado se hace un estudio previo o clasificación de los codificadores, para así mejor situar las contribuciones realizadas.

Los algoritmos de codificación de la voz se pueden dividir esencialmente en dos clases: algoritmos de codificación de la forma de onda y algoritmos de codificación paramétrica.

Los codificadores de la forma de onda analizan, codifican y reconstruyen la señal de entrada intentando en lo posible aproximar la evolución muestra a muestra de dicha señal. Para codificadores en el dominio del tiempo, la reducción en el *bit-rate* se realiza haciendo uso de la redundancia existente en la forma de onda. Tales redundancias se manifiestan en forma de periodicidades y dependencias estadísticas entre muestras sucesivas. A esta clase de algoritmos pertenecen las técnicas clásicas DPCM (*Differential PCM*) y ADPCM (*Adaptive DPCM*) [Jaya84]. En el dominio de la frecuencia se aprovecha la distribución no uniforme de la información de la voz en el espectro. Entre otros cabe citar los codificadores SBC (*Subband Coding*) [Jaya84] y ATC (*Adaptive Transform Coding*) [Zelins77] que obtienen buenos resultados para velocidades de transmisión superiores a 2 bits por muestra. Estos dos esquemas son la base para la nueva área de interés emergente recientemente de la banda ancha (voz o señales de audio de 7 kHz muestreadas a 16 kHz).

Por contra, en los algoritmos paramétricos, también denominados *vocoders* (de la contracción de los términos ingleses *voice coders*), en el decodificador no se intenta reproducir la forma de onda exacta de entrada sino solo una señal auditivamente equivalente a ella. Para ello usualmente se utiliza un modelo de producción de voz

en el que la señal es considerada ser el resultado de pasar una señal de excitación glotal a través de un filtro lineal variable en el tiempo que modela las características resonantes del tracto vocal. Lo que se codifica y envía a través del canal son los parámetros que caracterizan al modelo, recontruyéndose la señal de voz en el receptor a partir de dichos parámetros. Los parámetros decodificados no proporcionan información suficiente para regenerar la forma de onda de la señal de entrada, pero la información disponible debe ser suficiente para regenerar un sonido perceptualmente equivalente. A este tipo pertenecen los vocoders de predicción lineal (LPC vocoders) [Rab78] que operan usualmente a *bit-rates* inferiores a los usados en codificadores de la forma de onda, siendo valores típicos de funcionamiento 0.5 bits por muestra e inferiores. Esta reducción en *bit-rate* lleva una disminución de la naturalidad de la voz resultante aunque la inteligibilidad se mantiene razonablemente.

Además de los anteriores, debido al interés despertado en los últimos años, hay que citar explícitamente una categoría adicional, híbrida de los dos tipos anteriores. Como en el caso de los codificadores paramétricos, existe un modelo de producción de la señal de voz constituido por un filtro lineal variable con el tiempo, para el que se selecciona la excitación de forma tal que la señal sintetizada resultante sea lo más parecida posible a la forma de onda de la señal original, como en el caso de los codificadores de la forma de onda. Debido a que la selección de la secuencia de excitación depende de las diferencias entre la señal original y la sintetizada, estas técnicas requieren síntesis durante la fase de análisis, recibiendo por ello el nombre de *codificación usando predicción y análisis por síntesis.* De entre los codificadores híbridos cabe destacar los esquemas MPE (*Multipulse Excitation Coding*) [Atal82], actualmente en desuso, siendo más utilizado el esquema CELP (*Code Excited Linear Prediction*) [Atal85], que ha sido estudiado en profundidad en multitud de trabajos posteriores. Valores típicos de operación para estos esquemas están comprendidos entre 0.5 y 2 bits por muestra.

1.3 La Aproximación Propuesta

Para codificar los parámetros espectrales que caracterizan el filtro vocal, la mayor parte de las técnicas encontradas en la literatura intentan hacer uso de la correlación existente entre dichos parámetros en el dominio del tiempo. Esta correlación estadística se llamará *inter-trama.* A su vez, no hay que olvidar que, para un instante de tiempo dado, los coeficientes espectrales presentan una correlación estadística entre ellos, llamada *intra-trama.* Explotando pues estas dos correlaciones, se puede abordar el problema de la eliminación de la redundancia.

Para la zona de interés de muy bajo *bit-rate* son especialmente significativos los trabajos de cuantización matricial de Tsao y Gray [Tsao85], los vocoders por segmento de Roucos et al. [Rou82], [Rou83], de Shiraki y Honda [Shira88] [Honda92], así como el codificador multi-trama de Kemp et al [Kemp91], entre otros.

La aproximación adoptada consiste básicamente en reducir la redundancia inter-trama aproximando la evolución temporal de la trayectoria en el espacio de representación N-dimensional mediante la concatenación de N rectas. Toda la información a transmitir consistirá en caracterizar la longitud de las rectas, a la vez que el punto del espacio N-dimensional que la define (conocido el anterior). La determinación eficaz del mejor punto siguiente así como su localización temporal (mediante un procedimiento sencillo que se explicará en el capítulo 3) proporcionará el mejor compromiso rate-distorsión de entre todos los sistemas simulados para la zona de *bit-rates* de interés. En el procedimiento anterior, implícitamente se están utilizando las dependencias inter-trama, ya que los puntos N-dimensionales se codificarán mediante cuantización vectorial.

Independientemente del proceso anterior, para reproducir la forma de onda en el decodificador, es necesario disponer de una caracterización de la señal de excitación del filtro vocal. Dicha caracterización (suponiendo el modelo simplificado del proceso de generación de la voz), consiste por un lado, en la ganancia del filtro y por otro lado, en la frecuencia fundamental junto con la clasificación sonoro/no sonoro. En trabajos

anteriores [Wong82], [Lopez90c], se ha demostrado que una codificación diferencial de estos parámetros es suficiente. No obstante, dado que el proceso de codificación de la información espectral introduce un retardo significativo (varios retardos se estudian en [Lopez93a]), y por coherencia con la motivación principal del presente trabajo, se puede conseguir un mejor compromiso considerando las dependencias lineales existentes para dichos parámetros. En esta dirección se proponen algoritmos que hagan un uso eficiente de dichas dependencias. Los procedimientos utilizados serán la particularización monodimensional del algoritmo $N-dimensional$ propuesto para la información espectral.

Para todos los sistemas se proporcionarán los resultados experimentales obtenidos mediante simulación. Con ello, se podrá verificar su validez, así como cumplir el objetivo o motivación del presente trabajo, que no es más que responder a la pregunta: ¿dónde está el mejor compromiso experimental para la curva velocidad de transmisión frente a la distorsión para la fuente de señal de voz en la región de baja velocidad de transmisión o *bit-rate*, para retardos menores de 200 ms?.

1.4 Estructura del Trabajo Realizado

Para resolver los problemas planteados en la introducción de este capítulo, se ha llevado a cabo la investigación que se resume en esta memoria. Teniendo presente los objetivos ya establecidos y las aportaciones realizadas, el presente trabajo se ha estructurado de la siguiente manera. En este primer capítulo, se ha definido y acotado el problema a resolver, así como sus posibles aplicaciones. En el siguiente capítulo se presentan los fundamentos relativos al análisis y parametrización de la señal. Se estudia la posible adopción de un análisis en bucle cerrado, se resumen los estándares del Departamento de Defensa del Gobierno de los Estados Unidos FS-1015 y FS-1016. En el mismo capítulo, se fija y justifica la elección de las condiciones de análisis adoptadas. Para disponer de unas herramientas adecuadas al evaluar los esquemas elaborados, se ha dedicado el tercer capítulo al estudio de medidas de

calidad tanto objetivas como subjetivas. En dicho capítulo se justificarán las medidas de evaluación adoptadas. En el capítulo cuarto se resumen las aportaciones llevadas a cabo cuando se codifica la información espectral. En la medida de lo posible se han hecho comparaciones objetivas con sistemas desarrollados previamente en la bibliografía sobre la misma secuencia de test. En el siguiente capítulo se proponen varios esquemas para codificar la señal de excitación y se evalúa subjetivamente la unión de algunos de los esquemas propuestos para la codificación de la información de la envolvente espectral junto a esquemas para la codificación de la estructura fina. Como conclusión se finaliza con un capítulo donde se resume la investigación realizada.

Capítulo 2

La Señal de Voz: Análisis y Estándares de Codificación. Condiciones Experimentales

Antes de presentar los esquemas de codificación propiamente dichos, este capítulo se dedica a resumir los fundamentos relativos al análisis y parametrización de la señal. Se adoptará el modelo paramétrico de Predicción Lineal (LPC). Para la información espectral se presentarán distintos espacios de representación, eligiéndose por sus propiedades los parámetros LSP. Para la señal de excitación además de resumir los procedimientos clásicos de análisis, se estudia la posible adopción de un procedimiento de análisis en bucle cerrado. Por su proximidad en *bit-rate* al objetivo de esta tesis, se resumen los estándares FS-1015 y FS-1016 a 2400 y 4800 bps respectivamente. Finalmente se extractan las condiciones experimentales de análisis adoptadas en el resto del trabajo.

2.1 Un modelo para la producción de la voz.

Aun siendo de sobra conocido [Mark76], [Rab78], por razones de completitud, en este apartado se resumen brevemente las ideas básicas del modelo de producción de la voz, punto de partida esencial en el desarrollo de codificadores paramétricos e híbridos, y por tanto, utilizado en este trabajo de investigación.

Figura 2.1: Modelo de Producción de la Voz.

La producción de voz en el aparato fonador humano se puede modelar mediante una fuente que genera una señal de excitación que es filtrada por una serie de filtros en cascada que aproximan el modelado glotal, la cavidad bucal y los efectos de radiación de los labios, figura (2.1) Este modelo se puede describir matemáticamente en el dominio de la *transformada z* en la forma

$$S(z) = E(z)G(z)V(z)L(z) \tag{2.1}$$

donde $S(z)$, $E(z)$, $G(z)$, $V(z)$ y $L(z)$ son respectivamente la transformadas de la secuencia de salida, la excitación, la función de transferencia del filtro glotal, la función de transferencia del filtro del tracto vocal y el filtro de radiación de los labios.

Evidentemente, en la expresión (2.1) no se ha incluido explícitamente el tiempo como variable independiente, pero es fácilmente admisible que los filtros y la excitación deben actualizarse periódicamente para representar adecuadamente la naturaleza cambiante en el tiempo de la señal de voz. Así se habla de *análisis asíncrono* si dicha actualización se realiza a intervalos regulares de tiempo independientemente de la señal de entrada y *análisis síncrono* si se realiza de acuerdo con algunas características de la señal de entrada, como puede ser el periodo fundamental.

Usualmente, las contribuciones del filtro glotal $G(z)$ (que se modela con un filtro

todo-polos) se pueden considerar canceladas por el cero del filtro $L(z)$ de radiación [Mark76], quedando la expresión (2.1) reducida a

$$S(z) = \frac{E(z)}{A_p(z)} \tag{2.2}$$

donde el filtro $1/A_p(z)$ es un filtro todo-polos que modela las resonancias del tracto vocal y en general a la secuencia de filtros citada. $A_p(z)$ es también denominado el *filtro inverso.*

En este trabajo se adoptará un procedimiento de análisis asíncrono, con lo que implícitamente se está suponiendo que ambos procesos (la excitación y el posterior filtrado) están totalmente decorrelacionados, lo cual permite su análisis y codificación por separado. En el dominio de la frecuencia esta adopción permite analizar y codificar el espectro de la señal de voz en lo que se llama la *estructura fina* directamente relacionado con la señal de excitación y la *envolvente o información espectral* proviniente del filtro todo-polos considerado.

2.1.1 La Información Espectral. Predicción Lineal.

Para el análisis de la información espectral, una de las simplificaciones aceptadas es considerar el filtro inverso $A_p(z)$ cuasi-estacionario, es decir, suponer dicho filtro invariable durante periodos cortos de tiempo. Dicha suposición permite considerar los coeficientes de dicho filtro (expresión (2.3)) constantes durante los pequeños intervalos de análisis considerados, denominados *tramas.*

$$A_p(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{2.3}$$

Teniendo en cuenta las expresiones (2.3) y (2.2), la secuencia de excitación puede expresarse en el dominio del tiempo como la siguiente ecuación en diferencias

$$e(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{2.4}$$

que pone de manifiesto una de las implicaciones básicas del modelo considerado:

> *cada muestra de la señal de voz se puede predecir como una combinación lineal de las muestras anteriores*

En la bibliografía al análisis de la información espectral usando este modelo se le denomina *análisis de predicción lineal ó LPC ("Linear Predictive Coding").*

Definiendo $\hat{s}(n)$ como la muestra predicha en el instante n, se tiene que

$$e(n) = s(n) - \hat{s}(n) \tag{2.5}$$

Así, la secuencia de excitación $\{e(n)\}$ es también llamada *error de predicción*, siendo p el *orden de predicción* y a_i los coeficientes de predicción lineal.

Caracterizar o analizar la información espectral, consistirá pues en determinar exclusivamente los coeficientes del filtro a_i, con $i = 1, \cdots, p$. Dicho análisis será el resultado de la minimización de alguna función del error de predicción $\{e(n)\}$. En particular, debido a su fácil resolución matemática y fundamentalmente a los buenos resultados obtenidos, los coeficientes son determinados minimizando la energía de la secuencia del error de predicción α, definida por

$$\alpha = \sum_n e^2(n) \tag{2.6}$$

La minimización de la expresión anterior da lugar a un conjunto de p ecuaciones lineales con $\{a_i\}$ como incógnitas. Quedan por definir los límites de la sumatoria de la expresión (2.6); si la señal de entrada es considerada ser cero fuera del intervalo de la ventana de análisis, se llama *método de autocorrelación* y por el contrario se llama *método de covarianza* si la sumatoria es evaluada para $0 \leq n \leq N - 1$, siendo N la longitud de la ventana de análisis considerada. Teniendo presente las propiedades de diagonalización de la matriz asociada al anterior sistema de ecuaciones, en [Rab78] se proponen varios procedimientos recursivos para su resolución.

Parametrización de la Información Espectral

La información espectral resultado del análisis LPC, consiste en el conjunto de parámetros $\{a_i\}$, con $i = 1, \cdots, p$, que representan una estimación de la envolvente espectral para la *trama* analizada. Esta información puede representarse de múltiples formas equivalentes mediante transformaciones a otros espacios de representación o parametrizaciones que exhiban propiedades más adecuadas para su codificación. Baste decir, por ejemplo, que para un conjunto de coeficientes LPC dados, aun conteniendo toda la información de la envolvente espectral, no siempre estará garantizada la estabilidad del filtro asociado tras su codificación. Además es conocido el gran rango dinámico que presentan los coeficientes LPC, característica no deseable cuando dichos coeficientes van a ser cuantizados. Todo esto justifica la consideración de otros espacios de representación o parametrizaciones que aun siendo matemáticamente equivalentes a los LPC, tengan propiedades de interés para su codificación, como son:

- **Coeficientes PARCOR (PARtial CORrelation)** que para el método de autocorrelación pueden obtenerse con la recursión hacia atrás de la forma

$$k_i = a_i^{(i)}$$
$$a_j^{(i-1)} = \frac{a_j^{(i)} + a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2} \qquad \text{con} \qquad 1 \leq j \leq i-1 \tag{2.7}$$

  Donde el subíndice i va desde p hasta 1 e inicialmente se fija

$$a_j^{(p)} = a_j \tag{2.8}$$

  Los PARCOR tienen una potente propiedad para la estabilidad del filtro todopolos asociado, ya que es condición necesaria y suficiente que

$$-1 \leq k_i \leq 1 \tag{2.9}$$

para que los polos del filtro estén dentro del círculo unidad en el dominio de z. Los PARCOR son también llamados *Coeficientes de Reflexión*, porque se pueden interpretar como tales, entre las secciones adyacentes del modelo fisiológico de producción de la voz del *tubo acústico* [Mark76].

- **Coeficientes LAR** (Log Area Ratio), definidos por

$$g_i = \log\left(\frac{1-k_i}{1+k_i}\right) \qquad 1 \leq i \leq p \tag{2.10}$$

Que como se observa son una simple transformación bilineal logarítmica de los PARCOR. Los LAR están relacionados en el ámbito del modelo del tubo acústico con el logaritmo del cociente de las áreas de secciones adyacentes y tienen como propiedad interesante una sensibilidad espectral plana [Makh75], lo que se traduce en que se pueden cuantizar linealmente.

- **Parámetros LSP.** Más recientes que los anteriores, los coeficientes LSP *"Line Spectral Pairs"* se propusieron por F. Itakura y N. Sugamura en Japón, año 1975. Posteriormente se han desarrollado en el trabajo [Suga86]. Sus propiedades, analíticas y estadísticas, han sido estudiadas ampliamente en [Soong84], [Soong88] y [Soong90], donde también se proponen distintos codificadores. En el trabajo [Kang85] se propone y prueba un vocoder basado en los LSPs, a la vez que se hace un estudio completo y en profundidad de sus propiedades. También cabe citar los trabajos de N. Farvardin, donde se han propuesto con éxito varios codificadores aprovechando eficazmente las propiedades de estos parámetros junto a sus redundancias [Suga88], [Farvar89], [Pham90], [Laroia91] entre otros.

En general, los LSPs son por ahora (debido a las propiedades que a continuación se citan) el espacio de representación más idóneo para caracterizar la información contenida en el espectro LPC de la señal de voz.

Para un filtro inverso LPC, de orden p, dado por la ecuación (2.3), artificiosamente se puede extender el orden de dicho filtro a $(p+1)$ sin necesidad de introducir información adicional, sin más que hacer que el coeficiente de reflexión $(p+1)$–ésimo tome uno los valores ± 1. Esto es equivalente a dejar la última sección del tubo acústico totalmente cerrada o abierta.

Se puede demostrar [Rab78] que el filtro inverso $A_j(z)$ verifica la siguiente relación recursiva (donde el subíndice j indica genéricamente el orden de predicción)

$$A_j(z) = A_{j-1}(z) - k_j z^{-j} A_{j-1}(z^{-1}) \tag{2.11}$$

donde $A_0(z) = 1$ y k_j es el j-ésimo coeficiente PARCOR. En particular para $j = p+1$ se tiene que

$$A_{p+1}(z) = A_p(z) - k_{p+1} z^{-(p+1)} A_p(z^{-1}) \tag{2.12}$$

y para los valores artificiales y extremos de k_{p+1} elegidos, dicha recursión se puede expresar como

$$P(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \tag{2.13}$$

$$Q(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \tag{2.14}$$

donde se ha notado con $P(z)$ y $Q(z)$ a los dos nuevos polinomios obtenidos, que corresponden con cerrar o abrir totalmente la glotis en el modelo fisiológico del tubo acústico. Por construcción, se observa que $P(z)$ es un polinomio simétrico y $Q(z)$ es un polinomio anti-simétrico y que

$$A_p(z) = \frac{1}{2}[P(z) + Q(z)] \tag{2.15}$$

De la factorización de dichos polinomios, y si se considera p un número entero par, se tiene que [Suga88]

$$P(z) = (1 - z^{-1}) \prod_{i=2,4,\cdots,p} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{2.16}$$

y

$$Q(z) = (1 + z^{-1}) \prod_{i=1,3,\cdots,p-1} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{2.17}$$

donde se ha supuesto que $\omega_1 < \omega_3 < \cdots < \omega_{p-1}$ y que $\omega_2 < \omega_4 < \cdots < \omega_p$. De las expresiones (2.16) y (2.17) se puede comprobar que las raíces de dichos polinomios están en la circunferencia unidad, es decir son de la forma $e^{j\omega i}$ con $i = 1, 2, \cdots, p$.

Se define como parámetros LSP ("*Line Spectral Pairs*") al conjunto $\{\omega_i\}_{i=1,2,\cdots,p}$ de raíces de los polinomios $P(z)$ y $Q(z)$. Nótese que los valores de $\omega_0 = 0$ y $\omega_{p+1} = \pi$ son siempre raíces de $P(z)$ y $Q(z)$ respectivamente, que serán excluidas de la definición. Usualmente, en vez de la definición anterior, los coeficientes LSP se usan expresados en hertzios mediante la transformación $f_i = (\omega_i/2\pi)f_s$, donde f_s es la frecuencia de muestreo. Los LSP han sido interpretados físicamente como las frecuencias de resonancia de la cavidad bucal bajo las condiciones extremas artificiales de total apertura y clausura de la glotis [Suga88].

A continuación se enumeran las propiedades de los LSP que justifican su adopción como espacio de representación para la información espectral. La demostración de las mismas se encuentra en la bibliografía hasta aquí citada. En particular es de resaltar que

1. Todas las raíces de $P(z)$ y $Q(z)$ están en la circunferencia unidad.
2. Las raíces de $P(z)$ y $Q(z)$ verifican

$$0 = \omega_0 < \omega_1 < \omega_2 < \cdots < \omega_{p+1} = \pi \tag{2.18}$$

Siendo ω_i con $i = 2j$ $j = 0, \cdots, p/2$ y ω_i con $i = 2j+1$ $j = 0, \cdots, p/2$ raíces de $P(z)$ y $Q(z)$ respectivamente. Ésta es la *propiedad de la ordenación* que ha sido utilizada eficazmente para el diseño de un codificador para la información espectral en [Suga88]. Además, en la misma referencia, se establece que es condición necesaria y suficiente la verificación de esta propiedad para garantizar la estabilidad del filtro LPC asociado.

3. De la representación del espectro LPC y la localización de los parámetros LSP, se puede observar que la presencia de una frecuencia resonante (llamada *formante*) en el espectro, se corresponde con la localización de dos parámetros LSP muy próximos [Suga86]. En definitiva, el ancho de banda de los formantes es inversamente proporcional a las diferencias entre parámetros LSP adyacentes.

4. Para determinar cómo está distribuida la información entre los LSP, o mejor, cómo afectará individualmente el error cometido en cada LSP, se puede definir la *sensibilidad espectral* (SE_i) como

$$SE_i = \lim_{\delta \to 0} \frac{DS_i(\delta)}{\delta}, \qquad \text{con} \qquad i = 1, 2, \cdots, p \tag{2.19}$$

donde $DS_i(\delta)$ es la distorsión espectral (ver capítulo 3, apartado (3.3.3)) incurrida al modificar el coeficiente f_i en δ Hz. En la tabla (2.1) se muestran valores típicos [Suga88] de la sensibilidad espectral de cada uno de los parámetros LSP

Como se observa en la tabla (2.1) en general los primeros coeficientes tienen una SE mayor, de lo cual se concluye que caso de ser cuantizados escalarmente, deberían ser cuantizados mejor que los coeficientes de orden superior.

Coeficientes LSP	SE_i (dB/Hz)
ω_1	0.018
ω_2	0.016
ω_3	0.013
ω_4	0.012
ω_5	0.014
ω_6	0.012
ω_7	0.012
ω_8	0.012
ω_9	0.011
ω_{10}	0.012

Tabla 2.1: Sensibilidad Espectral Coeficientes LSP.

Comparación de los coeficientes PARCOR y LAR con los LSP

Como se ha dicho, los LSPs son actualmente el espacio de representación más idóneo para caracterizar la información espectral. Para reforzar esta idea en este subapartado se aportan algunos datos y conclusiones extraidas de estudios anteriores.

La preferencia de los LSPs frente a los PARCOR es corroborada en [Suga86], donde se comprueba experimentalmente que para obtener una distorsión espectral de 1 dB utilizando cuantización escalar uniforme para los PARCOR es necesario asignar 50 bits por trama, mientras que usando LSP, para obtener la misma distorsión sólo se necesitan 35 bits. Además en dicho trabajo también se demuestra que los LSP tienen unas características que los hacen más apropiados que los PARCOR cuando el *bit-rate* es reducido mediante interpolación. En particular, para la misma distorsión espectral el número de tramas codificadas es el 75% de las necesarias cuando se usan los PARCOR.

En [Soong84] se comparan objetivamente los LAR con la codificación de las diferencias entre LSP adyacentes, obteniéndose un 30% de reducción de *bit-rate* y una distorsión media de máxima semejanza (ver capítulo 3) de 0.069 frente a 0.097 obtenida con los LAR.

Además, dicha preferencia también ha sido evidenciada subjetivamente, por ejemplo en el informe presentado en [Kang85], donde al cuantizar independientemente los LSP de una trama con 33 bits, obtienen la misma puntuación DRT (ver apartado (3.2.1) que la obtenida con el estándar LPC a 2400 bps (ver apartado (2.3) de este capítulo) que usa 41 bits por trama para los PARCOR.

2.1.2 La Estructura Fina o Excitación

Habiendo supuesto que todo el modelado de la envolvente espectral se puede caracterizar por el filtro todo-polos $1/A_p(z)$ (expresión (2.3)), para concluir con el análisis de la señal de voz resta caracterizar la excitación a dicho filtro que determina la estructura fina del espectro.

El objetivo, en este caso, es caracterizar o parametrizar la señal de excitación al filtro $1/A_p(z)$, llamada *residuo* o error de predicción. Dicha caracterización se puede hacer tras un estudio detallado de la señal resultante de filtrar la señal original de entrada con el filtro inverso $A_p(z)$. En esta sección se distinguirá entre codificadores paramétricos e híbridos.

La Excitación en Codificadores Paramétricos

Para el tipo de codificadores llamados paramétricos o vocoders, tradicionalmente se realiza un clasificación grosera de las excitaciones atendiendo a su periodicidad. Suponiendo que las excitaciones deben tener una envolvente espectral plana que después será modelada por el filtro $A_p(z)$ y que tienen un carácter localmente estacionario, el modelo considerado adopta la siguiente simplificación

- la excitación es un tren de pulsos unitario con periodo inversamente proporcional a la frecuencia fundamental, llamados *tramas periódicas o sonoras*

- o se considera una excitación aleatoria, con una distribución gausiana blanca de varianza unidad, llamándose *tramas no periódicas o no sonoras.*

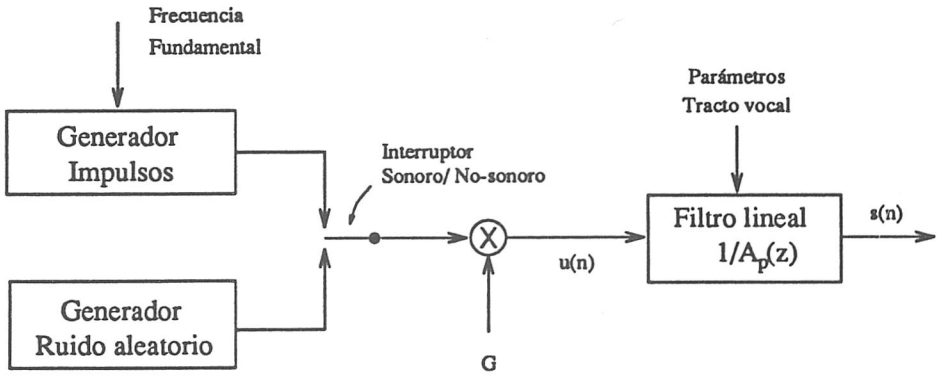

Figura 2.2: Modelo Simplificado de Producción de la Voz.

La señal de excitación así considerada es posteriormente multiplicada por un factor de ganancia G, como muestra la figura (2.2). Para caracterizar el modelo, mostrado en la mencionada figura, sólo es necesario codificar la información relativa a la clasificación periódico/no periódico, que suele incorporarse dentro de la información concerniente a la frecuencia fundamental, más el factor de ganancia G. Éste es el primer modelo en orden de complejidad y precisión crecientes, y ha sido utilizado tradicionalmente para la región de bajo y muy bajo *bit-rate* [Lopez90c].

Evidentemente la drástica simplificación realizada tiene como ventaja la sencillez del modelo considerado, lo que se traduce en un *bit-rate* reducido. Por contra, para este modelado hay que citar que:

1. No siempre las *tramas* pueden clasificarse claramente en sonoras o no sonoras. Es más, como la clasificación se hace fijando un umbral para el máximo normalizado de la función de autocorrelación, se pueden cometer errores, lo que implica la necesidad de un suavizado a posteriori de la decisión realizada

2. La parametrización realizada no está basada en ningún criterio de maximización de la calidad resultante

3. Del análisis de la señal residuo resultante de filtrar la entrada con el filtro inverso $A(z)$, en algunos casos es evidente que incluso tratándose de una trama sonora, el residuo dista mucho de ser un tren de pulsos equidistantes, es decir

en el modelado realizado se han hecho equivalentes el concepto de *sonoridad* con el concepto de *periodicidad*, cosa que no siempre es cierta

4. Además, observando la evolución temporal de la señal residuo es evidente que si se quiere modelar adecuadamente, es necesario considerar unos intervalos de análisis inferiores a los intervalos considerados para analizar la información espectral. Este hecho es coherente con el funcionamiento fisiológico de las cuerdas vocales y la glotis, que se modifican más frecuentemente que el tracto vocal.

La Excitación en Codificadores Híbridos

A diferencia del anterior análisis en bucle abierto realizado en los vocoders, es también clásico el llamado modelo de *predicción lineal excitado por código (CELP: Code-Excited Linear Prediction)* [Schroe85], perteneciente a los codificadores híbridos, donde típicamente para el análisis de la señal residuo se consideran intervalos mucho menores (del orden de 5 a 8 ms) y se realiza un procedimiento de análisis mediante síntesis en bucle cerrado (figura (2.3)). La idea básica consiste en, habiendo estimado el filtro que modela la envolvente espectral, encontrar la secuencia de excitación (de entre un conjunto previamente almacenado llamado *diccionario*) que minimice algún criterio de error entre la señal de entrada y la sintetizada. Es decir, en el codificador hay una réplica exacta del sintetizador usado en el decodificador.

En este caso la estructura fina del espectro es modelada mediante el índice correspondiente al elemento del diccionario elegido (con un criterio de fidelidad dado) y un filtro lineal recursivo variable con el tiempo $1/P(z)$ que es incorporado para modelar la periodicidad de las tramas sonoras. La expresión general para dicho filtro, (llamado filtro de *"pitch"* o de retardo largo) es de la forma

$$P(z) = 1 - \sum_{k=-(p-1)/2}^{(p-1)/2} b_k z^{-(M+k)} \qquad \text{con} \qquad p = 1, 3, \ldots \tag{2.20}$$

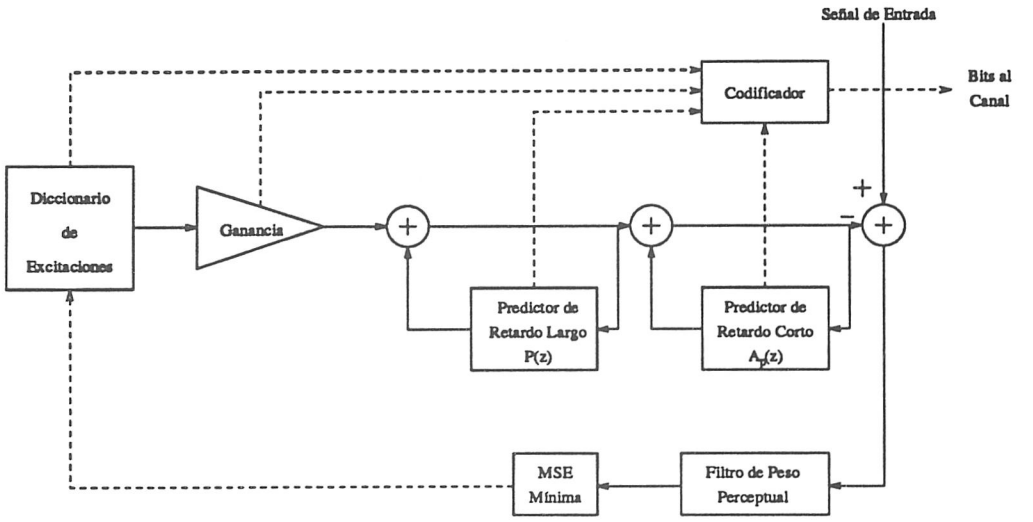

Figura 2.3: Diagrama por Bloques del CELP.

donde el valor de M es equivalente al periodo fundamental (o pitch) expresado en número de muestras (con valores típicos en el rango de 2 ms a 20 ms), y b_k son los coeficientes del filtro. Dichos coeficientes pueden ser obtenidos como siempre mediante un procedimiento de minimización del error de predicción, exactamente igual al realizado para la información espectral. Por lo tanto, a diferencia de los codificadores paramétricos, en este esquema no se hace una clasificación binaria y además la obtención de los parámetros se realiza maximizando un criterio de fidelidad.

La suposición básica para el diseño del diccionario consiste en admitir que, eliminando las periodicidades con el filtro de pitch, el residuo resultante se puede considerar que coresponde a una distribución gausiana de espectro plano.

Una aportación definitiva en el desarrollo de los codificadores de análisis mediante síntesis, fué la incorporación de un criterio perceptual en la minimización del error entre la señal sintetizada y la original. La idea básica consiste en distribuir el ruido existente, atendiendo a criterios de enmascaramiento propios del funcionamiento del

oído humano. Así se han propuesto pesos perceptuales [Atal79] de la forma

$$W(z) = \frac{A_p(z)}{A_p(z/\gamma)} \tag{2.21}$$

que distribuyen el ruido de cuantización de acuerdo con la propiedad de enmascaramiento. El parámetro γ controla la energía del ruido en los formantes, siendo un valor típico para 8 kHz de muestreo $\gamma = 0.90$.

La consideración de criterios perceptuales ha abierto una línea a considerar como determinante de lo que será la investigación en los próximos años en la codificación de la fuente [Jaya93].

El esquema CELP presentado, si bien puede considerarse una aproximación más acertada que el modelo simplificado de los codificadores paramétricos, no es del todo perfecta, ya que se está suponiendo que con el filtro $1/P(z)$ (típicamente de 3 coeficientes) se elimina toda la *periodicidad*, y lo que es más toda la *sonoridad*. Para soslayar este defecto, en la bibliografía se han propuesto soluciones alternativas tales como considerar filtros de pitch con retardos no enteros [Kroon91], o bien, eliminar dicho filtro, considerando la excitación como la suma de un vector código del diccionario estocástico más un vector código de un diccionario adaptable, construido con las excitaciones anteriores. Esta última idea ha sido utilizada en la elaboración del estándar FS-1016, resumido brevemente en una apartado posterior.

Otra contribución interesante puede encontrarse en [Shohan93], donde para los tramas sonoras se adopta una estrategia diferente, consistente en cuantizar vectorialmente la transformada de Fourier discreta del residuo (realizada síncronamente con el pitch), posteriormente interpolada en el receptor.

Análisis Paramétrico en Bucle Cerrado

Teniendo en cuenta todas las consideraciones de apartados anteriores e inspirados por los trabajos [Tzeng90] y [Tzeng91], aquí se propone un modelo de análisis alternativo para la señal residuo. El análisis a realizar se llevará a cabo adoptando algunas de las

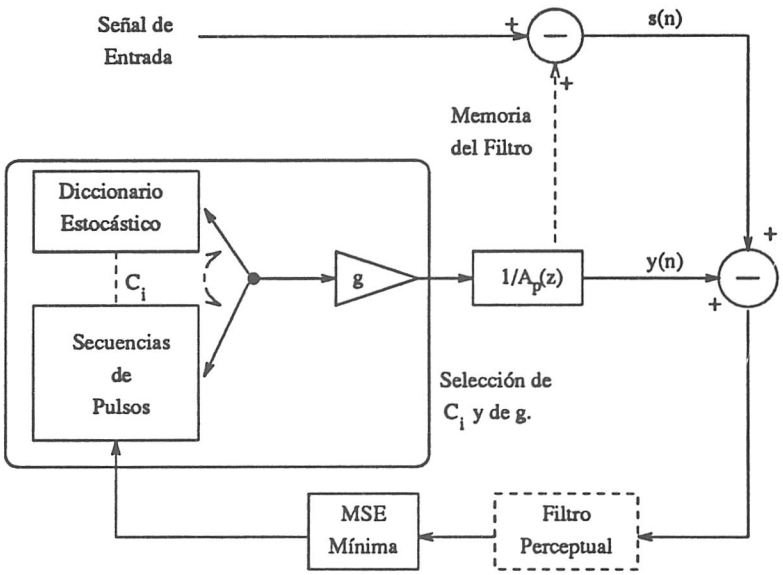

Figura 2.4: Análisis Paramétrico de la Excitación en Bucle Cerrado.

aportaciones positivas de los esquemas CELP, junto con la simplificación realizada en los codificadores paramétricos, necesaria siempre que se quiera reducir el *bit-rate*.

Para tal fin se propone el sistema representado en la figura (2.4), donde se realiza un análisis en bucle cerrado, pero considerando dos posibles modelos para la excitación, que será o bien un vector código correspondiente a un diccionario de secuencias gausianas (diccionario estocástico) o la excitación correspondiente a un tren de pulsos unitarios separados un número fijo de muestras, que corresponderá al periodo de pitch. Para el análisis de la excitación de cada trama se utilizará cada uno de los posibles C_i (figura 2.4) tanto del diccionario estocástico como cada una de las secuencias de pulsos correspondientes a los posibles periodos fundamentales considerados. Asímismo el factor de ganancia g se determina simultáneamente, minimizando el error cuadrático medio $E(g, C_i)$ mediante la expresión

$$(g, C_i) = \arg\min E(g, C_i) = \arg\min \sum_{n=1}^{N} [s(n) - gy(n)]^2 \qquad (2.22)$$

donde N es el número de muestras considerado para la trama de análisis, que en este

caso, es igual a la longitud considerada para el análisis de la información espectral. $s(n)$ representa a la señal de entrada, a la que se le ha substraido la memoria asociada al filtro inverso de retardo corto $1/A_p(z)$, y finalmente $y(n)$ representa la respuesta de dicho filtro a la excitación considerando el vector C_i.

Para un $y(n)$ dado, de la expresión (2.22), se obtiene que el valor de g que minimiza el error cuadrático medio es

$$g = \frac{\sum_{n=1}^{N} s(n)y(n)}{\sum_{n=1}^{N} y^2(n)} \tag{2.23}$$

Sustituyendo (2.23) en (2.22), minimizar $E(g, C_i)$ es equivalente a maximizar E'

$$E' = \frac{\left[\sum_{n=1}^{N} s(n)y(n)\right]^2}{\sum_{n=1}^{N} y^2(n)} \tag{2.24}$$

Las implicaciones de este análisis son

1. Para cada trama, hay una clasificación sonoro/no sonoro implícita, que se hará atendiendo a un criterio de fidelidad, ya que de entre todas las posibles excitaciones (secuencia gausiana o tren de pulsos), se eligirá aquélla que minimice el error cuadrático medio entre la señal de entrada y la sintetizada.

2. Caso de ser sonoro, el periodo de pitch obtenido obedecerá igualmente al mismo criterio de fidelidad

3. Al igual que en CELP se puede incorporar fácilmente en el análisis un criterio perceptual (figura (2.4))

Resultados Experimentales y Mejoras al Análisis Paramétrico en Bucle Cerrado

Al aplicar el análisis en bucle cerrado a toda la secuencia de test (apartado 3.5.2), considerando un diccionario estocástico de 256 vectores (8 bits), y un intervalo de periodos de pitch de 16 a 143 muestras (2.0 ms a 17.8 ms), al realizar la síntesis se

Diccionario Estocástico (vectores)	Separación entre pulsos (muestras)	SSNR (dB)
256	16 - 143	4.82

Tabla 2.2: SSNR en Bucle Cerrado.

obtuvo la relación señal ruido por segmentos (SSNR definida en la expresión 3.5) que se muestra en la tabla (2.2). Este pobre resultado es debido principalmente a que

1. Como ya se ha dicho se consideran tramas de 22.5 ms, valor excesivamente grande para modelar acertadamente la señal de excitación producida en las cuerdas vocales

2. Para las secuencias de pulsos se ha supuesto que el primer pulso siempre corresponde con la primera muestra de la trama, cosa que evidentemente no siempre es cierta, solo lo sería caso de hacer un análisis síncrono, en el que los impulsos de pitch estuvieran en fase con las ventanas de análisis. Este hecho, introduce errores en la clasificación sonoro/no sonoro, que suponen una disminución en la SSNR

Como mejoras al esquema de partida, teniendo en cuenta el punto 2 anterior, se incorpora la siguiente modificación al procedimiento anterior:

> *Si la trama anterior fué clasificada como no sonora y la actual es sonora, buscar el máximo valor absoluto de la señal en la trama actual y generar las secuencias de pulsos teniendo en cuenta que uno de los pulsos debe coincidir con la posición de dicho máximo.*

En definitiva, intentar hacer que la secuencia de impulsos esté en fase (o sincronizada) con la señal residuo. Incorporando esta idea, para la misma secuencia de test se obtuvieron los resultados mostrados en la tabla (2.3). El análisis cuyos resultados se

Diccionario Estocástico (vectores)	Separación entre pulsos (muestras)	SSNR (dB)
256	16 - 143	5.12

Tabla 2.3: SSNR en Bucle Cerrado Síncrono.

muestran en la tabla (2.3), aun mejorando la calidad objetiva, lleva implícitamente caso de ser codificado, un incremento en el *bit-rate*, ya que en este caso habría que transmitir información para caracterizar las diferencias de fase entre la ventana de análisis y el primer pulso de la secuencia de excitación. Ahora bien, esa información podría ser descartada y no codificada, en cuyo caso se tendría un vocoder donde la salida sería una señal que no minimiza el error cuadrático medio con la señal de entrada. Para estas condiciones, la SSNR deja de ser un criterio de evaluación significativo.

Además de las mejoras en SSNR, el aquí llamado análisis en bucle cerrado síncrono, introduce menos errores que el asíncrono, tanto en la estimación del periodo de pitch como en la clasificación sonoro/no sonoro.

Adicionalmente, intentando minimizar aun más los errores de clasificación sonoro/no sonoro y evitando a la vez la aparición de valores erróneos del periodo de pitch, se han incorporado las siguientes modificaciones al esquema anterior:

- Test de cruces por cero. Si el número de cruces por cero es mayor que 2 por ms la trama es clasificada no sonora.
- Imponer un umbral máximo para la ganancia g para los tramas no sonoras
- Imponer un umbral mínimo para la ganancia g para los tramas sonoras
- Favorecer la clasificación como sonoro al final de un tren de tramas sonoras
- Suavizar el pitch entre tramas adyacentes

- Realizar una interpolación lineal dentro de una trama sonora del factor de ganancia g

Finalmente, es de resaltar que si bien el número de errores al incorporar todas estas modificaciones evidentemente disminuye, la SSNR deja de ser una medida significativa, y lo que es más, tras varios test subjetivos informales, se puede concluir que comparado con un vocoder clásico (análisis en bucle abierto) la calidad obtenida en las zonas donde no hay error de clasificación es superior en el sistema desarrollado. No obstante, debido a que el análisis se realiza en bucle cerrado, un fallo en la clasificación de una trama afecta a las tramas posteriores.

2.2 El Modelo Sinusoidal

Aparte del modelo histórico del tubo acústico, y del modelo simplificado de producción de la voz (LPC), de todas las alternativas propuestas en la bibliografía, cabe citar el modelo sinusoidal, considerado como alternativa al esquema CELP en la región de 2.4-4.8 kbps.

La idea básica subyacente en todos los codificadores basados en un modelo sinusoidal consiste en sintetizar voz mediante la suma de términos seno, con amplitudes, frecuencias y fases variables en el tiempo. Estos codificadores pueden clasificarse dentro de los paramétricos pues no es su objetivo reproducir la forma de onda original.

Como trabajos significativos dentro de esta filosofía cabe citar los esquemas de *harmonic coding* [Marq90], *sinusoidal transform coding* [McAu92] y *multiband excitation coding* [Brands91].

Una de las características interesantes de estos esquemas, es que no sólo son capaces de modelar voz, sino que debido a la aproximación realizada de suma de senos, pueden codificar otro tipo de sonidos (música, ruido de fondo, etc) sin degradación de calidad.

Esta propiedad anterior junto con los resultados obtenidos hasta ahora, hacen

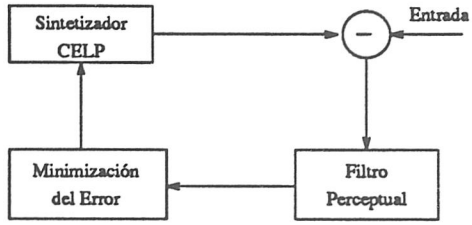

Figura 2.5: Análisis CELP.

del modelo sinusoidal una alternativa al ya clásico CELP, aunque dichos modelos no son necesariamente excluyentes [Tran90].

2.3 Estándares de Codificación

Sin ser un resumen exhaustivo del estado del arte de la codificación de la voz, (para un resumen completo ver [Gersho94]), en este apartado se comentan algunos de los estándares de codificación, que pueden considerarse como referencia obligada en la elaboración de este trabajo de investigación. La elección de los dos estándares aquí citados está hecha en base a los *bit-rates* de operación, en particular se extractan los dos estándares federales del gobierno de los Estados Unidos en la zona de muy bajo *bit-rate* FS-1016 a 4.8 kb/s y FS-1015 a 2.4 kb/s.

2.3.1 FS-1016 4.8 kb/s

Este estándar está basado en el esquema híbrido CELP de "análisis por síntesis". Las excitaciones al filtro asociado al predictor lineal son elegidas entre dos posibles diccionarios: uno de ellos estocástico pseudo-aleatorio con solapamiento (desplazamiento de -2 muestras) y trivaluado (-1, 0, +1) de 512 vectores código y el otro también con solapamiento, adaptable, que es actualizado por cada subtrama, con 256 vectores (figura 2.6).

Se usa ancho de banda telefónico (8 kHz de muestreo) y un tamaño de análisis para las tramas de 30 ms, divididas en 4 subtramas de 7.5 ms. El análisis en bucle cerrado, (figura 2.5) consiste en minimizar el error cuadrático medio pesado

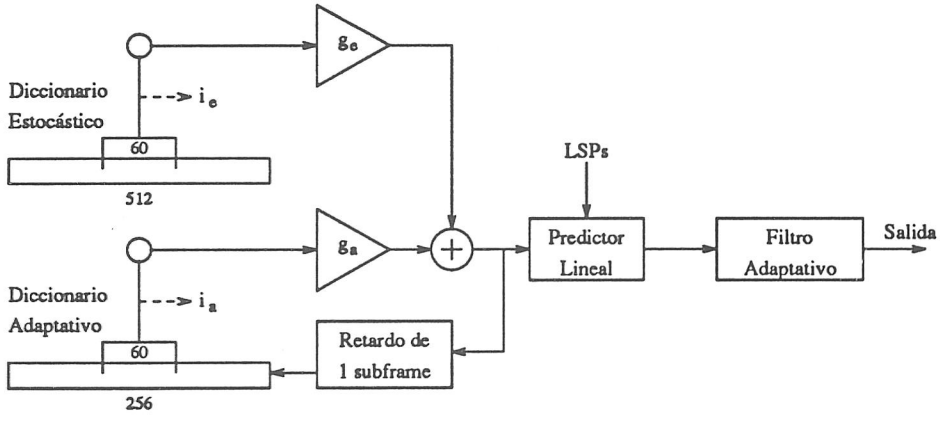

Figura 2.6: Síntesis CELP.

perceptualmente (ver [Kroon92]) entre la señal de entrada y la salida, realizando los siguientes pasos

1. determinar el filtro de predicción lineal (de retardo corto)
2. buscar la excitación en el diccionario adaptable (de retardo largo)
3. buscar en el diccionario estocástico

Para la síntesis, el procedimiento es esencialmente igual al análisis salvo que se incorpora un filtro de realce para mejor la calidad resultante, figura (2.6). En la tabla (2.4) se presenta un resumen de las características del FS-1016 a 4.8 kb/s [Camp91].

2.3.2 FS-1015 2.4 kb/s

Basado en el modelo simplificado para la producción de la voz, el estándar federal FS-1015 (también conocido por LPC-10) codifica la voz a un *bit-rate* de 2.4 kb/s. Recogido en el documento oficial [FS1015] y también descrito en [Trem82], ha sido un punto de referencia obligado en todo el desarrollo posterior de la codificación de la voz a muy bajo *bit-rate.*

El análisis, figura (2.7), se realiza tras un filtro de pre-énfasis (de la forma $H(z) = 1-0.9375z^{-1}$), por el método de la covarianza síncronamente con el periodo de pitch

	Predictor Lineal	Dicc. Adaptable	Dicc. Estocástico
Actualización	30 ms	30/4 = 7.5 ms	30/4 = 7.5 ms
Parámetros	10 LSPs	256 vectores	512 vectores
Análisis	lazo abierto correlación expansión BW 15 Hz Hamming 30 ms; no preénfasis	lazo cerrado 60 muestras peso = 0.8 rango = 20 a 147 búsqueda delta retardos no-enteros	lazo cerrado 60 muestras peso = 0.8 solapamiento -2 trivaluado
Bits/trama	34, LSP independientes {3,4,4,4,4,3,3,3,3,3}	índice: 8+6+8+6; ganancia(±): 5*4	índice: 9*4; ganancia (±): 5*4
bit-rate	1133.33 bps	1600 bps	1866.67 bps

Tabla 2.4: Características del Estándar FS-1016.

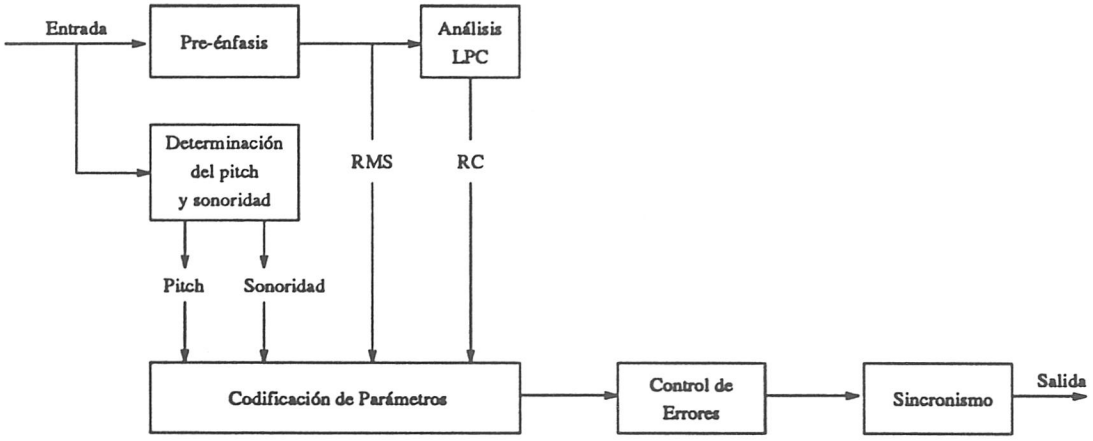

Figura 2.7: Análisis LPC-10.

para las tramas con sonoridad, y asíncronamente para las tramas no sonoras. El espacio de representación elegido para los coeficientes LPC es el de los coeficientes de reflexión. El periodo de pitch se determina en bucle abierto con el algoritmo AMDF (para un estudio comparativo de distintos algoritmos de estimación del periodo de pitch ver [Galvez91]). Se introduce un retardo de dos tramas para suavizar la caracterización sonora/no sonora de la trama así como el pitch. También se incluyen bits de redundancia para proteger la información contra ruidos en el canal además de un bit adicional para sincronización.

Posteriormente los parámetros son codificados de acuerdo con la asignación de bits mostrada en la tabla (2.5), para tramas de análisis de 22.5 ms de duración.

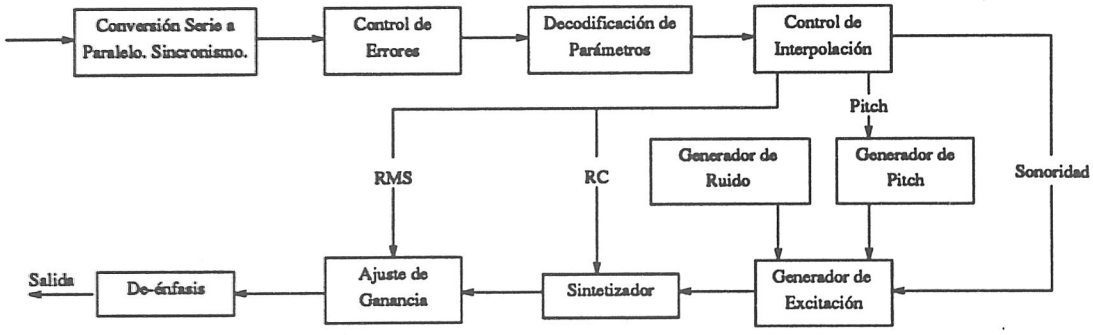

Figura 2.8: Síntesis LPC-10.

En la síntesis, figura (2.8) tras la conversión serie a paralelo, control de sincronismo, y control de errores, en el bloque de interpolación se introduce una trama de retardo para suavizar posibles errores. Posteriormente basándose en la caracterización de sonoridad se genera la excitación correspondiente, para después sintetizar realizando una conversión de los coeficientes de reflexión a coeficientes LPC del filtro.

El estándar fué modificado posteriormente [Kang82] y [Kang84], etiquetándose con el acrónimo ("LPC-10e"), siendo compatible con la especificación original.

Dado que las técnicas empleadas en el FS-1015 están claramente superadas por los recientes avances realizados por parte de la comunidad científica, es de resaltar, la elaboración de un nuevo estándar de alta calidad, actualmente en fase de proposición también a 2.4 kb/s, cuyos requerimientos básicos puede observarse en [Welch93]. La definición, elaboración y desarrollo de dicho nuevo estándar marcará en los próximos años la línea de investigación donde se emplearán los mayores esfuerzos para la codificación de la voz en la región de bajo *bit-rate* .

2.4 Condiciones Experimentales de Análisis

Para terminar este capítulo, se extractan las condiciones de análisis adoptadas en el resto del presente trabajo.

El ancho de banda considerado para la señal de entrada es de 4 kHz (ancho de

Coeficientes	Sonoros	No Sonoros
Pitch/Sonoridad	7	7
RMS	5	5
Sincronismo	1	1
k_1	5	5
k_2	5	5
k_3	5	5
k_4	5	5
k_5	4	0
k_6	4	0
k_7	4	0
k_8	4	0
k_9	3	0
k_{10}	2	
Suma	54	33
Corrección Errores	0	20

Tabla 2.5: Asignación de Bits en FS-1015.

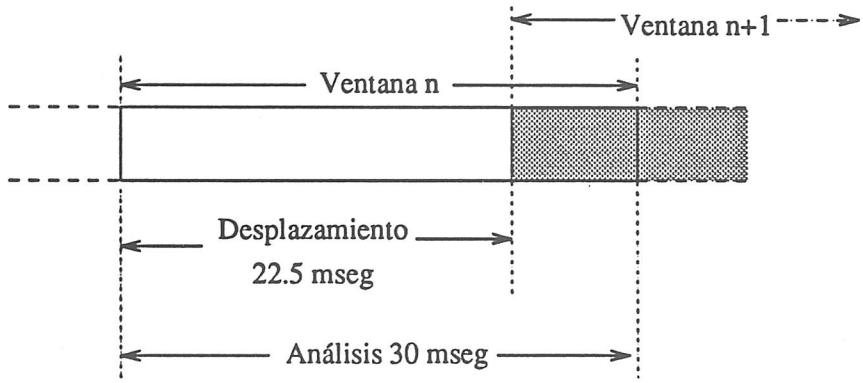

Figura 2.9: Longitud de la Ventana de Análisis.

banda telefónico), por lo que la frecuencia de muestreo (ó Nyquist) fué de 8 kHz. Se optó por el análisis asíncrono eligiendo una duración de las ventanas de análisis de 30 ms (240 muestras), con un desplazamiento de 22.5 ms para cada trama o ventana (ver figura 2.9).

Para el análisis de la información espectral (elegido el modelo simplificado, apartado (2.1)), no se realizó preénfasis, y se utilizó el método de autocorrelación para

Ancho de Banda	4 kHz
Frecuencia de Muestreo	8 kHz
Análisis LPC	Asíncrono, Autocorrelación
Ventana	Hamming
Duración ventana	30 ms
Desplazamiento	22.5 ms
Orden Predicción	10
Parametrización	Coeficientes LSP

Tabla 2.6: Condiciones de Análisis.

la obtención de los coeficientes del filtro LPC (expresión 2.3).

Adicionalmente, para una mejor estimación de los parámetros LPC, se utilizó una ventana de Hamming, definida por

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \leq n \leq N-1 \\ 0 & \text{en otro caso} \end{cases} \tag{2.25}$$

Siendo N el tamaño de la ventana de análisis.

El orden de predicción considerado fué $p = 10$, coherente con la regla empírica [Rab78] de que para modelar adecuadamente el espectro es necesario al menos 2 polos por kHz. Este orden de predicción podría aumentarse, pero es evidente que dicho aumento implicaría un aumento en la complejidad computacional asociada, por lo que la regla empírica antes citada es considerada ser un buen compromiso entre la resolución obtenida y el cálculo necesario.

En cuanto al espacio de representación, por todas las ventajas citadas en el apartado (2.1.1), se han elegido los coeficientes LSP.

Finalmente, en la tabla (2.6) se resumen todas las condiciones de análisis utilizadas en el desarrollo de este trabajo.

Capítulo 3

Procedimientos para la Evaluación de los Codificadores

Aun no siendo uno de los objetivos del presente trabajo, el propósito de este capítulo es estudiar algunos procedimientos de evaluación de los codificadores de voz, así como caracterizar su validez. Para lo cual, sin ánimo de exhaustividad, se resumen las técnicas más usuales encontradas en la bibliografía, usadas convencionalmente para la evaluación de sistemas de codificación- decodificación de voz. Algunas de los procedimientos presentados, serán utilizadas como herramientas básicas para la caracterización de los esquemas aportadas en este trabajo. Una clasificación natural de dichos procedimientos es la seguida en este capítulo, en el que se dedica un apartado a *Medidas Subjetivas* y otro a *Medidas Objetivas* de calidad. Finalmente, este capítulo concluye con un apartado dedicado a describir los corpus de test y entrenamiento usados para evaluar y diseñar todos los sistemas desarrollados.

3.1 Introducción

Es evidente que la tarea de la codificación no es más que la búsqueda del mejor compromiso entre la reducción de la cantidad de información empleada y la perdida de calidad ó distorsión introducida. Por tanto, para la comparación de distintos sistemas de codificación es necesario un procedimiento fiable y repetible que mida la calidad así como la cantidad de información necesaria para obtener esa calidad.

Si bien la cantidad de información para los sistemas de transmisión digitales es una magnitud fácilmente cuantificable (en bits por segundo), esta facilidad no lo es tal para la magnitud denominada *calidad.*

Cuando se habla del concepto de calidad hay dos cuestiones abiertas íntimamente relacionadas: cómo definir la calidad y cómo evaluar dicha calidad.

La definición rigurosa de la calidad de voz no es tarea fácil, ya que dicha definición está directamente relacionada con el proceso de percepción realizado en el ser humano. Es por tanto, una magnitud intrínsecamente subjetiva, que no puede ser rigurosamente definida.

Por esta razón, para evaluar a los sistemas de codificación se recurren a tests donde un conjunto de oyentes es encuestado sobre su opinión acerca de la calidad percibida para un conjunto de frases o bien, es forzado a que intente reproducir la secuencia de test, obteniéndose como medida el porcentaje de unidades perfectamente identificadas. Todas estas técnicas y otras que serán presentadas más adelante, donde se utiliza un conjunto de oyentes como procedimiento de evaluación, se denominan *subjetivas.* La evaluación subjetiva será un procedimiento valido, si no definitivo, para la determinación de la calidad, es decir, será un medida válida de la degradación introducida en el proceso de codificación-decodificación.

No obstante, los tests subjetivos presentan varios problemas, entre los cuales cabe citar por ejemplo lo costosos que son de realizar. Además de no ser fácilmente incorporables a un procedimiento de diseño iterativo, suelen ser resultado de una medida absoluta y lo que es aun peor, no son totalmente reproducibles por la inherente

no-reproducibilidad de las respuestas humanas.

Las razones anteriores justifican suficientemente la necesidad de otros procedimientos de evaluación, donde de una manera objetiva, se calcule la calidad de una forma más fiable, consistente, reproducible y menos costosa.

Es más, un procedimiento de medida de este tipo, no sólo servirá para la evaluación de sistemas de codificación sino que podrá ser parte de él, proporcionando un criterio de fidelidad que puede ser dinámicamente optimizado, tanto en la fase de diseño como en la de funcionamiento del codificador.

El problema evidente de las medidas objetivas, es que han de estar bien correlacionadas con el proceso de percepción, ya que en definitiva lo que se pretende es sustituir éste por aquéllas.

Mucho trabajo se ha llevado a cabo en este área, donde el objetivo final es determinar una medida objetiva que permita predecir la calidad subjetiva obtenible [Quack88], [Kita91], [Lopez90b], [T193]. Como resultado concluyente, todas las aportaciones se pueden resumir en que ninguna de las medidas objetivas propuestas puede sustituir para todas las situaciones (distintas tasas de transmisión) a las medidas subjetivas, y en definitiva al mecanismo de percepción. No obstante, para situaciones particulares, si que se han propuesto medidas que pueden ser y son de gran utilidad en el diseño, evaluación y funcionamiento de los codificadores, como las utilizadas en este trabajo.

3.2 Medidas Subjetivas de Calidad

En al apartado anterior se ha justificado la utilización de medidas objetivas de calidad. No obstante, es obvio que la medida de la *calidad de la voz* debe estar basada en el proceso de percepción. Además, la validez de un test objetivo debe determinarse en relación a cómo de predecible es la calidad subjetiva conseguible para una medida objetiva dada.

En este apartado se presentan varios procedimientos de evaluación subjetiva,

que se clasifican atendiendo a dos factores de calidad [Kita91]: *inteligibilidad y naturalidad.* Hay que notar que considerando las técnicas actuales de codificación la evaluación de la *naturalidad* es particularmente útil para velocidades de transmisión altas, mientras que los tests de *inteligibilidad* son especialmente útiles cuando se trata de codificadores a bajas velocidades de transmisión. Además de los factores de naturaleza monodimensional antes indicados, al final de este apartado se cita una medida de carácter multidimensional, llamada medida de aceptabilidad, en la que varias características se evalúan independientemente.

3.2.1 Métodos para determinar la *Inteligibilidad*

Los test de inteligibilidad son apropiados para aquellos sistemas que proporcionan la señal de voz con calidad significativamente degradada, donde los tests de opinión pueden proporcionar unos resultados poco fiables.

Medidas de Articulación

Las mediadas de articulación se llevan a cabo cuando se quiere evaluar la capacidad de transmisión de información del codificador considerado.

Dado un corpus de referencia, se define la *articulación* como el porcentaje de unidades de voz correctamente identificadas por parte de un conjunto de oyentes tras la codificación. Las unidades a identificar pueden ser fonemas, sílabas, palabras o frases. Es evidente que los niveles semántico y sintáctico de la información transmitida afectarán a dicho porcentaje. En otras palabras, siempre es de esperar una articulación superior cuando se consideran frases como unidades a identificar, frente a fonemas. Por ejemplo, en [Jaya84] se muestra que para un valor del 50% de aciertos, tomando como unidades un conjunto de consonantes, se consigue un 80% de aciertos considerando frases como unidades a identificar.

"Rhyme Tests"

Si se quiere evaluar la articulación de sonidos, eliminando la interacción de los niveles semántico y sintáctico, es necesario la utilización de un corpus de palabras aisladas carentes de sentido. Es, por tanto, usual utilizar un conjuntos de 50 monosílabas en la forma consonante-vocal-consonante, donde el oyente debe identificar la primera consonante, conociendo la vocal y la última consonante.

Para facilitar la realización de este tipo de medidas, se limitan las opciones a elegir por el oyente a un conjunto máximo de 6 opciones para cada una de las palabras del test [Kita91].

Una modificación a los anteriores es el DRT "Diagnostic Rhyme Test" en el que se impone que las opciones presentadas al oyente, no sólo se diferencien en la primera consonante, sino que además difieran en sólo una de las características distintivas de dicha consonante. Para el DRT, las opciones ofrecidas al oyente se reducen a dos, siendo sólo una de ellas la correcta.

Las características distintivas utilizadas son: sonoridad, nasalidad, sustentación, sibilación, gravedad y compacidad. Una clasificación de los fonemas consonantes ingleses según éstas características puede encontrase en la referencia [Quack88].

La principal ventaja del DRT es que no sólo proporciona una medida de la inteligibilidad para un codificador dado, sino que además proporciona un diagnóstico para cada una de las características citadas anteriormente.

Los porcentajes de inteligibilidad varían típicamente entre el 70 y el 95 por ciento [Dimo93]. Para codificadores de alta calidad las puntuaciones suelen estar en torno al 90 por ciento, y como desventaja del DRT, cabe citar que para este tipo de codificadores es difícil obtener resultados estadísticos suficientemente significativos, lo que justifica la utilización de otras medidas basadas más en la *naturalidad* que en la *inteligibilidad.*

Puntuación	Calidad	Distorsión
5	Excelente	Imperceptible
4	Buena	Perceptible pero no molesta
3	Aceptable	Perceptible y moderadamente molesta
2	Pobre	Molesta pero no rechazable
1	Mala	Muy molesta y rechazable

Tabla 3.1: Puntuaciones MOS.

3.2.2 Métodos para determinar la *Naturalidad*

Para un codificador con calidad suficiente, es decir, cuando la inteligibilidad es alta, todavía es posible apreciar características que subjetivamente hagan que el oyente prefiera o considere mejor la calidad de un sistema frente a otro. Todas estas características se engloban en el concepto denominado *naturalidad.* En este apartado se presentan distintos procedimientos para evaluar subjetivamente la naturalidad.

Tests de Opinión Media (MOS)

El test MOS ("Mean Opinion Score") es el procedimiento más utilizado para la evaluación subjetiva de sistemas de codificación. En este método, los oyentes puntúan la calidad de la señal sintetizada, usando la escala mostrada en la tabla (3.1), basándose en su propia impresión. La calidad o el valor obtenido con este test, será la media aritmética de todas las puntuaciones proporcionadas por el conjunto de oyentes, sobre el mismo corpus.

Entre las ventajas de este test cabe citar que con el procedimiento seguido, la evaluación realizada es global, en el sentido de que todas las características de la señal son evaluadas conjuntamente. Por contra como inconveniente cabe citar la poca reproducibilidad de los resultados obtenidos. Como nota ilustrativa es conveniente mencionar el experimento realizado en [Kita84], donde para unas mismas condiciones experimentales, mismo corpus y usando los mismos codificadores, se pueden observar diferencias de más de 1 punto en los MOS obtenidos, dependiendo sólo del pais donde

el MOS fué realizado.

Para soslayar esta deficiencia, el test MOS requiere de una fase de entrenamiento de los oyentes, donde se les entrena o se proporcionan algunas señales de referencia junto con su puntuación asignada a priori.

Validez de los tests MOS

Como ya se ha dicho, uno de los problemas asociados con las medidas subjetivas en general y el test MOS en particular, es la carencia de reproducibilidad de los resultados obtenidos. Factores que influyen en la variabilidad de la opinión por parte del oyente son: la capacidad auditiva, raza, sexo, lugar de origen, actitud emocional durante el test y por supuesto tener criterios diferentes de interpretación de calidad.

Es por tanto conveniente, cuando se ofrecen resultados MOS, medir la variabilidad de los oyentes, estimando la varianza:

$$\sigma_L^2 = \frac{1}{n-1} \sum_{q=1}^{5} n_q (q - MOS)^2, \tag{3.1}$$

donde n es el número total de oyentes y n_q es el número de oyentes que dieron una puntuación q. Valores usuales para σ_L son 0.6 a 0.8 [T193]

Tests de Comparación de Pares

Cuando la calidad a evaluar es alta, usando un MOS los resultados obtenidos se concentraran entre los valores 4 y 5. Para estos casos es conveniente evaluar los codificadores de una manera más afinada, donde al oyente se le presentan dos frases y es forzado a elegir entre uno de los dos estímulos ofrecidos. El resultado del test se ofrece como el porcentaje de preferencias para cada una de las opciones.

La ventaja de estos tests es que son más precisos que los MOS, sin embargo tienen el problema de que a veces es difícil determinar cual de los dos estímulos es peor si ambos tienen un tipo de degradación distinta, con lo que cuando la calidades

son próximas, el resultado puede no ser suficientemente significativo.

3.2.3 Medidas de *Aceptabilidad*

Al contrario que las anteriores, los tests de *aceptabilidad* (DAM "Diagnostic Acceptability Measure"), son una medida multidimensional de la calidad, en los que la voz es evaluada en varias escalas separadas, puntuadas independientemente. Los oyentes son invitados a evaluar un corpus para el que deben dar una puntuación para cada una de las 21 características previstas [Panzer93]. 10 de las características están relacionadas con la calidad perceptual de la señal, 8 están relacionadas con el sonido de fondo y las 3 restantes evalúan la inteligibilidad, la agradabilidad y la aceptabilidad general. Finalmente, estas puntuaciones son combinadas con unos pesos experimentales, para proporcionar una medida unidimensional.

Como puede intuirse, es éste uno de los procedimientos más específicos y precisos, pero tiene el inconveniente de ser muy costoso y de realización muy elaborada.

3.3 Medidas Objetivas de Calidad

Todos los procedimientos de evaluación subjetivos hasta ahora propuestos, presentan una complejidad alta a la vez que tienen un coste elevado. Es necesario proporcionar procedimientos de evaluación directamente calculables, sin la intervención de un conjunto de oyentes, y que puedan realizarse de una manera automática.

En general, como se muestra en la figura (3.1) las medidas objetivas se calculan a partir de un conjunto de señales de entrada E (no degradadas) y un conjunto de señales de salida S, resultado de la codificación-decodificación de E.

Una medida objetiva será tanto mejor cuanto más correlacionada esté con las medidas subjetivas. Además debe ser expresable analíticamente, de modo que sea fácilmente incorporable en el diseño y funcionamiento del codificador.

En el proceso de percepción intervienen todos los elementos del lenguaje, es decir el oyente responde al contenido contextual, semántico, prosódico, sintáctico y

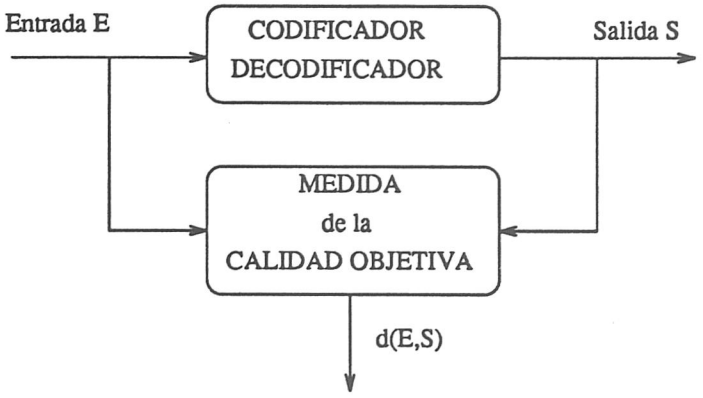

Figura 3.1: Cálculo de Medidas Objetivas.

fonético del mensaje. Un error en cualquiera de estos elementos se puede traducir en una pérdida total de la inteligibilidad del mensaje. Desde este punto de vista, una medida objetiva, que no incorporara todos los elementos del lenguaje no podría predecir la calidad subjetiva adecuadmente. Sin embargo, ya que los codificadores utilizados sólo introducen distorsión en el nivel acústico, puede concluirse que una medida objetiva, aun considerando sólo elementos acústicos, puede predecir adecuadamente la calidad subjetiva conseguible.

Dependiendo del dominio de operación, las medidas objetivas se pueden clasificar en medidas de la distorsión de la forma de onda (cuando se estiman en el dominio del tiempo) ó medidas de distorsión espectrales (cuando se opera en el dominio de la frecuencia). Para estas últimas cabe aun un refinamiento adicional, considerando por un lado aquellas que usan la estructura fina del espectro frente a otras que sólo consideran la envolvente espectral.

3.3.1 Validez de las Medidas Objetivas

Para la elección de una de entre las medidas objetivas que se van a proponer, es necesario determinar su validez. Un procedimiento adecuado para estimar la validez de una medida objetiva es determinar su capacidad de predecir la calidad subjetiva.

Es ésta una tarea prolija pues necesita disponer de grandes bases de datos, así

como la consideración de suficientes esquemas de codificación (para comprobar la validez de la medida en cuestión), además de disponer de una medida subjetiva de referencia que sea reproducible y fiable.

El análisis de regresión lineal puede ser una herramienta útil para medir la validez de una medida objetiva [Dimo89], para lo cual es necesario estimar el coeficiente de correlación entre las medida objetiva y subjetiva dado por

$$\rho = \frac{\sum_d (S_d - \bar{S}_d)(O_d - \bar{O}_d)}{\left[\sum_d (S_d - \bar{S}_d)^2\right]^{1/2} \left[\sum_d (O_d - \bar{O}_d)^2\right]^{1/2}} \tag{3.2}$$

done d es un índice para las medidas obtenidas, S_d es la medida subjetiva, O_d es la correspondiente medida objetiva, y la barra indica el valor medio estimado para todo d.

Se ha determinado que para que la medida objetiva pueda sustituir con fiabilidad a la opinión de 20 oyentes, se debe verificar que $\rho^2 \geq 0.95$, [T193].

Otra medida interesante y significativa para estimar la validez de una medida objetiva es la varianza del error cometido cuando la medida subjetiva es reemplazada por la medida objetiva, definida por

$$\sigma_e^2 = \sigma_s^2(1 - \rho^2) \tag{3.3}$$

siendo σ_s la desviación estándar de las medidas subjetivas. En definitiva esta expresión muestra qué valor de ρ^2 es necesario para conseguir una calidad objetiva con σ_e^2, cuando se tiene σ_s.

3.3.2 Medidas de Distorsión de la Forma de Onda

Las medidas objetivas en el dominio del tiempo miden directamente la distorsión entre la señal de entrada y la salida. A continuación se citan las medidas más utilizadas.

Relación Señal Ruido (SNR)

Aunque la SNR se usa convencionalmente para la caracterización de líneas de transmisión, también se ha usado como un parámetro objetivo para la evaluación de la calidad de voz. Puede ser calculada según la expresión

$$SNR = 10 \log \frac{\sum_{i=1}^{N} x(i)^2}{\sum_{i=1}^{N} [y(i-m)/A - x(i)]^2} \tag{3.4}$$

donde $x(i)$ y $y(i)$ son la i-ésimas muestras en la entrada y en la salida, respectivamente, N es el número de muestras, la constante A es la ganancia del codificador y finalmente m es el retardo introducido por el sistema. Es evidente que para una medida acertada es necesario introducir una estimación del parámetro A de normalización así como el factor de retardo m. Baste citar por ejemplo que si no fuese así, multiplicar la señal de entrada por el factor -1 introduciría una $SNR = -6\,dB$.

Relación Señal Ruido por Segmentos(SSNR)

Aunque la SNR es una buena medida de la degradación introducida en sistemas analógicos de comunicación, en [Quack88] se han mostrado sus deficiencias al evaluar sistemas digitales. Fundamentalmente, sus carencias están relacionadas con el hecho de que los intervalos de mayor potencia pesan más que los intervalos de potencia menor. Como alternativa más adecuada que la anterior, es común el uso de la $SSNR$, que no es más que el promedio de la SNR evaluada en intervalos limitados de tiempo denominados segmentos.

$$SSNR = \frac{1}{M} \sum_{j=1}^{M} SNR_j \tag{3.5}$$

siendo M el número de segmentos y SNR_j la relación señal ruido del segmento j. Valores típicos para la duración de los segmentos considerados varían entre 16 y 32 ms. Las ventajas se la $SSNR$ sobre la SNR residen en que esta medida pesa por igual a todos los intervalos independientemente de su potencia.

Para esta medida es conveniente evaluar sólo los segmentos que no corresponda a intervalos de silencio, ya que para estos, una pequeña cantidad de ruido se traduce en una pobre $SSNR$.

Se han propuesto otras variantes de la relación señal ruido, por ejemplo la SNR pesada en frecuencias, y la $SSNR$ granular, ver la referencia [Quack88], intentado obtener una mayor correlación con la calidad subjetiva asociada.

Es evidente que este tipo de medidas son significativas exclusivamente cuando el codificador tiene como premisa de diseño aproximar lo mejor posible las formas de onda entre la entrada y la salida, sin embargo, cuando el codificador se diseña para aproximar los espectros, estas medidas de distorsión no son adecuadas.

3.3.3 Medidas de Distorsión Espectral

En este apartado se van a considerar medidas en el dominio de la frecuencia considerando la estructura fina de los espectros de entrada y salida.

Distorsión Espectral

Se define como

$$DE = \left[\frac{1}{w}\int_0^w |S_e(w) - S_s(w)|^2, dw\right]^{1/2} \tag{3.6}$$

donde $S_e(w)$ y $S_s(w)$ son respectivamente los espectros de entrada y salida, que pueden ser calculados usando la Transformada de Fourier. Al igual que para la SNR, se suele calcular la Distorsión Espectral promedio, evaluada para segmentos con duración típica de 16 a 32 ms.

Esta medida de distorsión se usa con frecuencia para caracterizar la bondad de un codificador espectral [Suga88], en particular es comúnmente aceptado [Laroia91] que un codificador espectral puede considerarse *transparente* si introduce un $DE \leq 1\,dB$.

Una modificación a la anterior, es considerar un pesado espectral para distintas

bandas de frecuencias, obteniéndose la WDE ("Weighted Spectral Distortion")

$$WDE = \frac{1}{B}\sum_{b=1}^{B} W_b DE_b \tag{3.7}$$

siendo DE_b la distorsión espectral para la subbanda b, B el número de subbandas consideradas y W_b el coeficiente de peso. Los coeficientes de peso se estiman teniendo en cuenta características perceptuales, para con ello darle más importancia a las zonas del espectro más relevantes subjetivamente.

Medidas basadas en Función de Coherencia ("SDR")

Para esta medida se utiliza la correlación cruzada entre los espectros de entrada y de salida, definiendo la función de coherencia como

$$\hat{\gamma}^2(f) = \frac{\left|\sum_{n=1}^{N} X_n(f)Y_n^*(f)\right|^2}{\sum_{n=1}^{N}|X_n(f)|^2 \sum_{n=1}^{N}|Y_n(f)|^2} \tag{3.8}$$

donde $X_n(f)$ e $Y_n(f)$ son los espectros de entrada y salida, e $Y_n^*(f)$ es el complejo conjugado de $Y_n(f)$. La función $\hat{\gamma}^2(f)$ puede interpretarse como un coeficiente de correlación definido para cada frecuencia f.

A partir de la expresión anterior, se define la Relación Señal Distorsión SDR como el cociente entre la función de coherencia y la no-coherencia $(1 - \hat{\gamma}^2(f))$ expresada en decibelios como

$$SDR(f)_{dB} = 10\log\frac{\hat{\gamma}^2(f)}{1-\hat{\gamma}^2(f)} \tag{3.9}$$

Es ésta una de las medidas objetivas candidatas a ser estandarizadas por el Grupo de Estudio (SG) XII de la CCITT (ahora UIT). En [T193] se presenta una definición más detallada de la función de coherencia, sus propiedades y expresiones de regresión para la estimación de la puntuación MOS a partir de ella. Asimismo, se muestran resultados de la validez de la medida propuesta, obteniéndose para una de las bases

de datos consideradas un factor $\rho = 0.92$, definido en la expresión (3.2).

3.3.4 Medidas de Distorsión de la Envolvente Espectral

Al igual que las anteriores, este tipo de medidas operan en el domino de la frecuencia. Ahora bien, los espectros de entrada y salida son considerados más groseramente, tomando sólo su envolvente espectral.

La justificación de esta aproximación reside en que el oído, actua como un analizador de Fourier grosero. En particular, como se corroborará en este trabajo, la inteligibilidad está básicamente contenida en la envolvente espectral, luego tomando una medida que considere sólo la envolvente se estará aproximando mejor el comportamiento del oído, con lo que cabe esperar una mayor correlación de este tipo de medidas con la percepción subjetiva.

Medidas L_p basadas en el Análisis LPC

Uno de los métodos para obtener la envolvente espectral está basado en el análisis de Predicción Lineal (LPC), y consiste en calcular la respuesta en frecuencias del filtro todo-polos LPC. Con ello se obtendrá una versión suavizada del espectro asociado al intervalo de análisis ó trama considerada.

Dicho lo anterior, se puede concluir que cualquier medida basada en calcular diferencias entre los parámetros que caracterizan al filtro LPC entre la entrada y la salida, será en definitiva una medida de distorsión de la envolvente espectral. Parámetros usuales en la caracterización del filtro todo-polos (ver apartado 2.1.1) son los coeficientes de predicción, los coeficientes PARCOR ("PARtial CORrelation"), los coeficientes LAR ("Log Area Ratio"), [Mark76] ó los coeficientes LSP ("Line Spectral Pairs"), también llamados LSF ("Line Spectral Frequencies") [Kang85].

Para este tipo de representaciones se usan como medida de la distorsión las

distancias L_p, definidas por la expresión:

$$L_p = \frac{1}{N} \cdot \sum_{j=1}^{N} \left[\frac{1}{m} \cdot \sum_{i=1}^{m} |y_j(i) - x_j(i)|^p \right]^{\frac{1}{p}} \tag{3.10}$$

donde $x_j(i)$ e $y_j(i)$ son los i-ésimos parámetros de entrada y salida respectivamente del segmento j. Lo usual es considerar distancias Euclídeas, es decir tomar $p = 2$. También se puede optar por aproximar las diferencias entre los logaritmos de las envolventes espectrales, usando la expresión:

$$d_p = \frac{1}{N} \cdot \sum_{j=1}^{N} \left[\frac{1}{m} \cdot \sum_{i=1}^{m} |20 \log |y_j(i)/x_j(i)||^p \right]^{\frac{1}{p}} \tag{3.11}$$

Medida de la Razón de Semejanza Logarítmica

Suponiendo que la voz puede representarse como un modelo autoregresivo todo-polos, esta medida de distancia está relacionada con la disimilitud existente entre los modelos asociados a la señal de entrada y a la de salida. Se define la razón de semejanza logarítmica por

$$d(\vec{a}_e, \vec{a}_s, j) = \log \left(\frac{\vec{a}_s \mathbf{R}_e \vec{a}_s^T}{\vec{a}_e \mathbf{R}_e \vec{a}_e^T} \right) \tag{3.12}$$

donde j indica la posición del segmento, $\vec{a}_e$ es el vector de coeficientes LPC de entrada, $\vec{a}_s$ es el vector de coeficientes LPC de salida. $\vec{a}^T$ indica el vector traspuesto de $\vec{a}$ y finalmente $\mathbf{R}_e$ es la matriz de autocorrelación de la señal de entrada. Esta medida puede ser mejor interpretada considerando los filtros inversos asociados. En particular, sea $A_e(z)$ el filtro inverso que modela la envolvente espectral del segmento de entrada $x_e(n)$,

$$A_e(z) = 1 - \sum_{i=1}^{p} a_e(i) z^{-i} \tag{3.13}$$

Igualmente se puede definir el filtro inverso $A_s(z)$ para el segmento de salida. La medida de la Razón de Semejanza Logarítmica, se puede calcular como el logaritmo del

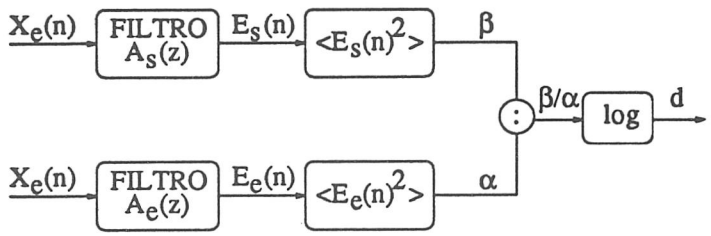

Figura 3.2: Interpretación de la Razón de Semejanza.

cociente entre las potencias de las señales residuo α y β, resultantes de filtrar inverso la forma de onda del vector de entrada con los filtros $A_e(z)$ y $A_s(z)$ respectivamente, como se indica en la figura (3.2). Es decir,

$$d(\vec{a}_e, \vec{a}_s, j) = \log(\beta/\alpha) \tag{3.14}$$

siendo

$$\beta = r_x(0)r_a(0) + 2\sum_{i=1}^{M} r_x(i)r_a(i) \tag{3.15}$$

donde $r_x(i)$ y $r_a(i)$ son las secuencias de autocorrelación de la señal de entrada $x_e(n)$ y de los coeficientes de $A_s(z)$ respectivamente.

$d(\vec{a}_e, \vec{a}_s, j)$ es una medida no-negativa ya que por definición del análisis LPC, siempre $\beta \geq \alpha$, y la igualdad sólo se cumple cuando los dos filtros $A(z)$ son iguales, es decir, cuando no hay distorsión.

Para la razón de semejanza (también llamada distancia de Itakura) es aceptado [T193] que a partir de valores que verifiquen que $d(\vec{a}_e, \vec{a}_s, j) < 1.4$, la degradación introducida no se considera significativa.

La razón de semejanza logarítmica, y sus variantes (distancia Itakura, distancia Itakura-Saito, distancia Itakura-Saito optimizada en ganancia, distancia Itakura-Saito normalizada en ganancia, etc), han sido estudiadas y utilizadas ampliamente en la bibliografía, como por ejemplo en los trabajos [Gray80], [Shore83], [Juang84], [Lopez90b]. Asimismo, distintas interpretaciones analíticas de las medidas propuestas pueden encontrarse en [Juang82].

Distancia Cepstral

En este apartado se presenta otra medida de la distorsión basada en estimar diferencias entre las envolventes espectrales. En concreto, se opera en el dominio de los coeficientes cepstrum [Oppen75], resultantes de un análisis homomórfico de la señal de voz. El cepstrum se obtiene calculando la transformada inversa de Fourier del logaritmo del espectro de potencias del segmento de voz. Lo usual, en vez de usar la estructura fina del espectro, es usar la envolvente espectral obtenida a partir del análisis LPC, en particular se puede establecer la siguiente relación de recursión [Rab78] para la obtención de los coeficientes cepstrales $C(n)$ en función de los coeficientes de predicción $a(n)$.

$$C(n) = a(n) + \sum_{i=1}^{n-1}(i/n)C(i)a(n-i) \qquad \text{para} \qquad n > 0 \tag{3.16}$$

Aunque la secuencia de coeficientes cepstrales es infinita, en la práctica es suficiente considerar tantos como el orden de predicción empleado en el análisis LPC.

Los coeficientes cepstrales se han usado con éxito tanto en aplicaciones de reconocimiento de voz [Rab85], [Peina94], [Segu91] (entre otras) como para la verificación e identificación de locutores, como por ejemplo en [Furui81].

Para la evaluación de la calidad, la distancia cepstral (DC) se puede definir como

$$DC = \frac{10}{\log_e 10}\left[2 \cdot \sum_{i=1}^{P}(C_e(i) - C_s(i))^2\right]^{1/2} \tag{3.17}$$

donde $C_e(i)$ y $C_s(i)$ son los i-ésimos coeficientes cepstrales derivados del análisis LPC, expresión (3.16). Hay otras expresiones similares para el cálculo de distancias en el cepstrum, pero en este trabajo se adopta la expresión anterior, por ser en particular la explicitamente considerada por el grupo de estudio (SG) XII de la CCITT como candidata al estándar para la evaluación de sistemas de procesamiento de voz [T193]. En [Kita88] se muestra un ejemplo de la reproducibilidad de la distancia cepstral, para la que se obtiene un coeficiente de correlación (3.2) con la puntuación MOS

$\rho = 0.97$, así como la expresión de la regresión cuadrática entre ambas medidas.

3.4 Medidas de Calidad Utilizadas

Una vez resumidas las técnicas de evaluación más significativas usadas para el procesamiento de la voz, en este apartado se enumeran y justifican, de entre las presentadas, cuales de ellas serán utilizadas en este trabajo.

3.4.1 Medidas Utilizadas para la Evaluación Objetiva de la Calidad

Dado que ninguna de las medidas objetivas presentadas puede sustituir definitivamente al proceso de la percepción, en vez de utilizar una sola de las medidas citadas, en este trabajo se considerarán las siguientes:

- *Distorsión Espectral.* Experimentalmente se ha demostrado que medidas definidas en el dominio de la frecuencia, como la distorsión espectral, expresión (3.6), reflejan mejor la calidad subjetiva que aquellas definidas en el dominio del tiempo. Además, esta medida ha sido recientemente utilizada en la bibliografía para evaluar los codificadores a baja razón de bits por segundo, y es obviamente útil para medir la degradación introducida en sistemas que codifican la información espectral en vez de caracterizar la evolución temporal de la señal.

- *Distancia Itakura-Saito normalizada en ganancias.* Si bien las medias que operan sobre la estructura fina del espectro son útiles para evaluar la calidad, es conocido que el procedimiento de la percepción es más sensible a la envolvente espectral que a la estructura fina. En particular, cuando se trata de codificadores a muy baja velocidad de transmisión, es más significativo intentar medir objetivamente la *inteligibilidad* que no la *naturalidad.*

  Las medidas de estimación de la razón de semejanza son útiles para este cometido. En particular, debido a su invarianza con la ganancia del codificador,

se usará una de las variantes de la Razón de Semejanza Logarítmica, conocida por la distorsión de Itakura-Saito normalizada en ganancias ("DISNG") [Shore83], definida por

$$DISNG(\vec{a}_e, \vec{a}_s) = \beta/\alpha - 1 = \frac{r_x(0)}{\alpha} r_a(0) + \left(2 \sum_{i=1}^{M} \frac{r_x(i)}{\alpha} r_a(i) \right) - 1 \quad (3.18)$$

Donde se ha utilizado la misma notación introducida en la expresión (3.15).

En ([Lopez90b]) se ha demostrado experimentalmente la particular utilidad de esta distancia frente a otras cuando se usa cuantización vectorial.

- *Distancia Cepstral.* Aunque este trabajo se centra en muy baja razón de bits por segundo, además de las anteriores, por completitud, se evaluarán los codificadores usando la distancia cepstral dada en la expresión (3.17), de especial utilidad para estimar la *naturalidad.* En particular, esta distancia se ha demostrado ([Kita82]) estar bien correlacionada con la medida subjetiva de opinión MOS.

- *Relación Señal Ruido.* La utilización de esta medida, expresión (3.4), está indicada para codificadores en los que la premisa de diseño esté en aproximar tanto como se pueda la forma de onda de entrada. Si bien no está demostrada una correlación clara entre esta medida y otras de naturaleza subjetiva, se utilizará porque a falta de otro criterio mejor, la SNR es una medida clásica para los codificadores citados. En particular, esta medida se utilizará para evaluar los codificadores escalares propuestos para la cuantización de los parámetros característicos de la estructura fina del espectro.

3.4.2 Medidas Utilizadas para la Evaluación Subjetiva de la Calidad

Debido a los bajos *bit-rates* de interés para este trabajo de investigación, se ha utilizado como medida de calidad subjetiva un test de inteligibilidad *"rhyme test"* donde para cada uno de los sistemas a evaluar se tomaron 25 sílabas (consonante-

vocal-consonante) extraidas aleatoriamente de locutores masculinos y femeninos no pertenecientes a la secuencia de entrenamiento.

Para cada una de las sílabas, los observadores del test fueron invitados a elegir una de entre seis posibilidades conociendo la vocal y consonante final. La *inteligibilidad* se estimó como el número de aciertos promediado para todas las sílabas y los 8 observadores encuestados.

3.5 Corpus de entrenamiento y test

En este apartado se detallan las particularidades de las bases de datos de voz usadas tanto para el entrenamiento como para realizar los tests. Dada la dificultad y lo costoso de la construcción y diseño de una base en castellano que recoja de forma suficientemente significativa toda la variedad existente a todos los niveles del idioma, se ha optado por la utilización de parte de la base de datos norteamericana TIMIT.

3.5.1 Secuencia de entrenamiento

Para la secuencia de entrenamiento de los sistemas para la codificación de la información espectral, se han considerado 80.000 ventanas de voz decimadas con un factor 1:2 (la TIMIT original está muestreada a 16k muestras por segundo) , cada una de ellas con duración de 22.5 ms (en total 30 minutos), extraidas de forma uniformemente distribuida entre las 8 regiones dialectales consideradas en el diseño de la TIMIT. La proporción de oyentes masculinos es del 70% frente al 30% de locutores femeninos, respetando las proporciones establecidas en el corpus total original. Todo este material ha sido elegido del subcorpus de entrenamiento sugerido en la TIMIT.

A su vez, para el entrenamiento de los cuantizadores de la señal de excitación se ha considerado un subconjunto aun más reducido, consistente en 45517 ventanas de la misma longitud que para la información espectral, y que es denominada TIMIT-2. Igualmente se han recogido las 8 variedades dialectales consideradas, y por cada

región o variedad dialectal se han elegido tres locutores masculinos y tres femeninos.

3.5.2 Secuencia de test

Siguiendo las sugerencias recomendadas en la TIMIT, para el test, se ha elegido el subcorpus denominado *"core test"* que contiene 24 locutores, 2 masculinos y 1 femenino por cada una de las 8 regiones dialectales consideradas. En total 192 frases, que constituyen un total de 22366 ventanas de voz correspondientes a más de 8 minutos de test. Este subcorpus se ha comprobado tener al menos una aparición de cada uno de los fonemas ingleses. Además, ningún locutor de los pertenecientes al "core test" está presente en la secuencia de entrenamiento por la TIMIT. Luego el no solapamiento está garantizado entre las secuencias de test y entrenamiento.

Este mismo corpus se utilizó para evaluar objetivamente los cuantizadores de la estructura fina o señal de excitación.

Para evitar un sobreentrenamiento del silencio y una importancia excesiva en la secuencia de test, se realizó un recorte de los silencios tanto al principio de cada frase como al final, de acuerdo con la segmentación sugerida en la TIMIT.

Capítulo 4

Codificación de la Información Espectral

En este capítulo se aportan una serie de algoritmos con la intención de responder a la pregunta formulada en la introducción de esta memoria: ¿Cómo aproximar experimentalmente las curvas $R(D)$ o $D(R)$ para la información espectral de la señal de voz?

Para lo cual, tras formular explícitamente el problema en la introducción, y después de caracterizar la fuente considerada, se plantean soluciones bien conocidas como son la Cuantización Escalar Óptima, la Cuantización Vectorial y algunas de sus variantes, su generalización a más dimensiones, y finalmente se propone un algoritmo donde se combina eficazmente un procedimiento de interpolación lineal con el proceso de cuantización. Este último será un procedimiento que de una manera sencilla y realizable responda a la pregunta anteriormente formulada.

4.1 Introducción

Uno de los objetivos en la codificación es eliminar toda la redundancia existente en la señal generada por la fuente de una forma recuperable, es decir, sin perdida de información en el receptor.

Dicho en otras palabras, el diseño de un codificador consiste en proporcionar un algoritmo ó procedimiento cuyo *bit-rate* y distorsión se sitúen sobre un punto de la curva $R(D)$, expresión (1.1).

Suponiendo un modelo estadístico conocido para la fuente y para algunos casos particulares, la *"Teoría Rate-Distorsión"* [Berger71], rama de la *Teoría de la Información*, puede proporcionar una expresión analítica para dichas curvas. Ahora bien, aunque se han desarrollado algoritmos iterativos y probadamente convergentes para la obtención de la curva $R(D)$ para un caso general [Blaut72] donde la función densidad de probabilidad de la fuente es conocida, dichos formalismos no proporcionan ningún procedimiento sistemático para diseñar el codificador que aproxime la curva $R(D)$ ó $D(R)$, que toda fuente tiene.

Además, la caracterización de algunas fuentes, como la señal de voz, por un modelo estadístico sencillo, o sea, una función densidad de probabilidad tratable matemáticamente no es una tarea trivial. En la bibliografía se han propuesto algunas aproximaciones para el modelado estadístico de la señal de voz [Jaya84], si bien ninguna de ellas se puede considerar aceptable para desarrollar un tratamiento riguroso y teórico del problema.

Debido a estas razones, en este capítulo, se aporta una solución experimental al problema planteado, donde el objetivo es proporcionar un algoritmo que aproxime las citadas curvas. Mientras que en situaciones prácticas, cuestiones tales como la complejidad, retardo introducido y *bit-rate* fijo son de importancia, sería deseable experimentalmente conocer cual es el compromiso límite alcanzable si tales restricciones no estuvieran presentes. Esencialmente éste es el problema estudiado en este capítulo.

En el desarrollo realizado se considerará sólo y exclusivamente la información espectral correspondiente al modelo autoregresivo de producción de la voz por el tracto vocal, y en particular, como se justificó en el apartado (2.1.1), se considerarán sólo los parámetros LSP, si bien gran parte de los procedimientos desarrollados podrían trasladarse a otras representaciones paramétricas del filtro LPC.

4.2 Autocovarianzas Normalizadas *Intra-trama* e *Inter-trama* de los LSPs

Antes de seguir adelante, en este apartado se realizará una caracterización estadística de la fuente considerada.

Una medida utilizada en la bibliografía, como indicador de la máxima redundancia teórica que un codificador puede eliminar [Mark76], es la SFM *"Spectral Flatness Measure"*, que puede aproximarse como el cociente entre la media geométrica y la media aritmética del espectro de densidad de potencias de la señal [Jaya84]. la SFM es una cantidad no negativa, y es igual a 1 si y sólo si el espectro es plano. Cuanto más plana sea la función densidad de potencia espectral, menos predecible será la fuente, luego menos redundancia exhibirá.

No obstante aquí, por sencillez, se caracterizará la fuente estimando las autocovarianzas normalizadas existentes entre los coeficientes LSP dentro de un mismo vector, llamada *covarianza intra-trama* dada por la matriz $\Phi = [\phi_{i,j}]$, tabla (4.1), con elementos $[\phi_{i,j}]$, $i, j = 1, 2, \cdots, p$ calculados como

$$\phi_{i,j} = \begin{cases} \dfrac{\sum_{m=0}^{M-1} (\omega_{m,i} - \bar{\omega}_i)(\omega_{m,j} - \bar{\omega}_i)}{\sum_{m=0}^{M-1} (\omega_{m,i} - \bar{\omega}_i)^2} & : \quad j \leq i \\ \phi_{j,i} & : \quad j > i \end{cases} \tag{4.1}$$

donde se ha usado $[\omega_{n,1}, \omega_{n,2}, \cdots, \omega_{n,p}]$ para notar a los coeficientes LSP correspon-

i/j	1	2	3	4	5	6	7	8	9	10
1	1.00	0.72	0.016	-0.24	-0.35	-0.58	-0.54	-0.63	-0.54	-0.45
2	0.72	1.00	0.82	0.58	0.26	0.01	0.01	0.02	-0.02	-0.09
3	0.01	0.82	1.00	0.93	0.56	0.52	0.38	0.45	0.27	0.12
4	-0.24	0.58	0.93	1.00	0.69	0.56	0.43	0.38	0.29	0.11
5	-0.35	0.26	0.56	0.69	1.00	0.76	0.45	0.37	0.20	0.09
6	-0.58	0.01	0.52	0.56	0.76	1.00	0.62	0.51	0.29	0.11
7	-0.54	0.01	0.38	0.43	0.45	0.62	1.00	0.72	0.46	0.23
8	-0.63	0.02	0.45	0.38	0.37	0.51	0.72	1.00	0.51	0.27
9	-0.54	-0.02	0.27	0.29	0.20	0.29	0.46	0.51	1.00	0.41
10	-0.45	-0.09	0.12	0.11	0.09	0.11	0.23	0.27	0.41	1.00

Tabla 4.1: Coeficientes $\phi_{i,j}$ de Autocovarianza *Intra-Trama*.

dientes a la trama n-ésima, siendo M el número de tramas consideradas y como siempre la barra indica el valor medio de la variable.

Como puede observarse en la tabla (4.1), hay una dependencia lineal bastante significativa entre los LSPs correspondientes a una misma trama, lo que justifica la utilización de algún procedimiento de eliminación de la redundancia existente al realizar la codificación.

Igualmente, se puede estimar la autocovarianza existente entre coeficientes pertenecientes a tramas adyacentes, que se llama *covarianza inter-trama*, dada por la matriz $\Psi = [\psi_{i,k}]$, tabla (4.2), con elementos $[\psi_{i,k}]$, $i, k = 1, 2, \cdots, p$ calculados como

$$\psi_{i,k} = \frac{\sum_{m=0}^{M-k-1} (\omega_{m,i} - \bar{\omega}_i)(\omega_{(m+k),i} - \bar{\omega}_i)}{\sum_{m=0}^{M-1} (\omega_{m,i} - \bar{\omega}_i)^2} \tag{4.2}$$

De la fuerte dependencia lineal observable en las tablas (4.1) y (4.2), se puede adelantar que un codificador que pretenda aproximar la curva $R(D)$, deberá utilizar las dependencias intra-trama e inter-trama. Y no solo eso, sino que además deberá aprovechar eficazmente las dependencias no lineales que exhiba la fuente, como se

i/k	1	2	3	4	5	6	7	8	9	10
1	0.81	0.69	0.61	0.56	0.52	0.49	0.47	0.47	0.47	0.46
2	0.72	0.51	0.38	0.31	0.27	0.25	0.23	0.23	0.22	0.21
3	0.77	0.57	0.43	0.35	0.30	0.28	0.28	0.29	0.29	0.29
4	0.82	0.67	0.55	0.46	0.42	0.39	0.38	0.38	0.38	0.38
5	0.86	0.72	0.58	0.47	0.39	0.34	0.30	0.28	0.26	0.25
6	0.84	0.68	0.53	0.41	0.33	0.27	0.24	0.22	0.22	0.21
7	0.82	0.65	0.51	0.39	0.30	0.24	0.21	0.20	0.20	0.20
8	0.81	0.64	0.49	0.37	0.30	0.25	0.23	0.22	0.22	0.23
9	0.75	0.57	0.43	0.33	0.27	0.24	0.23	0.22	0.22	0.22
10	0.73	0.56	0.44	0.36	0.31	0.28	0.27	0.27	0.26	0.25

Tabla 4.2: Coeficientes $\psi_{i,k}$ de Autocovarianza *Inter-Trama.*

verá en los apartados siguientes de este capítulo.

4.3 Cuantización Escalar Óptima

Como se ha ilustrado en la tabla (4.1), la fuente considerada exhibe una dependencia intra-trama significativa. Para hacer uso de esta dependencia lineal, se ha propuesto [Rou82] el siguiente codificador escalar óptimo, en el que se realizan los siguientes tres pasos:

1. *Decorrelación de los parámetros*, usando una matriz de transformación adecuada (por ejemplo la transformada óptima de *Karhunen-Loéve* [Jaya84] ó aproximaciones subóptimas para el caso general como la DCT *Transformada Discreta del Coseno* entre otras), obtener unos parámetros decorrelacionados en el dominio de la transformada.

2. *Asignación óptima de bits* para cada uno de los parámetros en el nuevo dominio.

3. *Cuantización Escalar*, codificar cada uno de los parámetros en el dominio de la transformada usando tantos bits como los asignados en el paso anterior con un simple procedimiento Lloyd-Max [Lloyd82].

De esta manera, las dependencias lineales mostradas se pueden eliminar consiguiéndose un buen compromiso R-D.

4.4 Cuantización Vectorial (VQ)

A pesar de que la cuantización óptima escalar elimina las correlaciones mostradas en la tabla (4.1), se han propuesto generalizaciones del procedimiento de Lloyd-Max a más de una dimensión, dando lugar a los llamados *Cuantizadores Vectoriales (VQ).* Dicha generalización está justificada por los siguientes hechos:

1. *Generalización de la $R(D)$ a fuentes con memoria.* Inicialmente se definió la R(D) como la mínima información mutua promedio conseguible para todos los posibles asignaciones de los símbolos de entrada con los símbolos de salida para una fuente dada, expresión (1.1). Imponiendo ciertas condiciones para la fuente [Berger73], una definición alternativa y consistente con la anterior, propuesta por Shannon, es

$$R(D) = \lim_{N \to \infty} R_N(D) \tag{4.3}$$

donde con $R_N(D)$ se hace referencia a considerar como símbolos de entrada N muestras consecutivas formando un vector N-dimensional. Por la propia definición, es obvio que en el límite teórico la $R(D)$ se puede aproximar tanto como se quiera usando un VQ con vectores de tamaño $N \to \infty$.

2. *Existencia de dependencias no-lineales.* Además de las autocovarianzas mostradas anteriormente, la fuente considerada exhibe otro tipo de dependencias no-lineales que no son utilizables por un codificador escalar óptimo. Experimentalmente se ha demostrado [Rou82],[Makh85] cómo un VQ puede conseguir un mejor compromiso que un codificador escalar óptimo, y ese compromiso es tanto mejor cuanto mayor es la dependencia estadística no-lineal en la fuente.

3. *Utilización de medidas de distorsión multidimensionales.* Dada la naturaleza multidimensional de la fuente considerada, una ventaja adicional de los VQ frente a los escalares es la posibilidad de usar medidas de distorsión multidimensionales como la Distorsión Itakura-Saito Normalizada en Ganancias, expresión(3.18).

Por las razones esgrimidas en los puntos anteriores, en vez de cuantizar cada parámetro separadamente, se considerarán varios formando un vector $p-$dimensional $\vec{x} = [x_1, x_2, \cdots, x_p]$, que será cuantizado como una unidad. Cuantizar vectorialmente $\vec{x}$ no es más que asignar un vector $p-$dimensional $\vec{y}$ a $\vec{x}$, de entre un conjunto finito $\mathrm{Y} = \{\vec{y}_i, 1 \leq i \leq L\}$ de cardinal L de acuerdo con un operador de cuantización $\mathbf{q}(.)$

$$\vec{y} = \mathbf{q}(\vec{x}) \tag{4.4}$$

Al conjunto $\mathbf{Y}$ se le denomina diccionario, L es el tamaño del diccionario e $\{\vec{y}_i\}$ es el conjunto de vectores código.

En la definición anterior del cuantizador vectorial $\mathbf{q}(.)$ hay implícita una partición $\mathcal{P} = \{C_i; i = 1, \cdots, L\}$ del espacio $p-$dimensional en un conjunto de L regiones o celdas C_i. Cada región tiene un representante ó vector código $\vec{y}_i$.

Para transmitir la información relativa a $\vec{x}$, tras realizar la cuantización usando el operador $\mathbf{q}(.)$, se enviará un índice binario c_i de longitud l_i bits, que identifica a la celda C_i a la que pertenece $\vec{x}$. En el receptor o decodificador, habiéndose recibido c_i, se generará como salida el correspondiente vector código $\vec{y}_i$. Es evidente que el número de bits l_i asignados a cada celda dependerá de la probabilidad a priori de que el vector de entrada sea cuantizado usando dicha celda, en definitiva de la estadística de la fuente. La velocidad de transmisión expresada en bits por segundo, en el caso general será una magnitud variable, que dependerá del vector de entrada y de la función de distribución de la fuente. En general, el *bit-rate* medio por vector

R se puede estimar como

$$R = \sum_{i=1}^{L} P(C_i) l_i \tag{4.5}$$

donde $P(C_i)$ es la probabilidad de que la fuente genere un vector perteneciente a la celda C_i.

Para el caso particular que haga una asignación de bits uniforme (que será la óptima para el caso de que la fuente y el cuantizador $\mathbf{q}(.)$ hagan todas las celdas equiprobables) se tendrá que

$$l_i = l = \log_2 L \qquad 1 \leq i \leq L \tag{4.6}$$

4.4.1 Diseño del Cuantizador Vectorial

Dado un número fijo de niveles (vectores) de cuantización L, el objetivo principal del diseño consiste en seleccionar dichos niveles junto con la elección de una partición del espacio de representación en regiones o celdas tal que se consiga la distorsión promedio menor posible. En concreto, se trata de determinar los vectores $\vec{y}_i$ y las regiones C_i con $i = 1, \cdots, L$, tal que se minimice

$$D = \sum_{i=1}^{L} \int_{C_i} (\vec{x} - \vec{y}_i)^2 f_{\vec{X}}(\vec{x}) dx \tag{4.7}$$

donde $f_{\vec{X}}(\vec{x})$ representa la función densidad de probabilidad de $\vec{X}$.

El problema así planteado no tiene una solución directa; ahora bien, descomponiendo el diseño del cuantizador en dos problemas, que no son otros sino considerar por un lado el codificador y por otro el decodificador, se proporciona una estrategia intuitiva y eficaz para aproximar el diseño del cuantizador óptimo.

Fijada una de las dos partes consideradas, es fácil determinar las condiciones necesarias que se han de verificar para optimizar la otra. Por tanto, se deben cumplir dos condiciones, que no son más que la generalización multidimensional del procedimiento monodimensional de Lloyd-Max:

1. *Regla de mínima distorsión o vecino más próximo*, procedente de determinar el codificador óptimo fijado un decodificador. Intuitivamente es obvio que si el objetivo es minimizar la distorsión, ningún codificador será mejor que aquél que elija el vector código $\vec{y}_i$ que minimice la distorsión al vector de entrada $\vec{x}$. Formalmente, dado un diccionario con vectores y_i, las celdas deben ser tales que

$$C_i = \{\vec{x} : d(\vec{x}, \vec{y}_i) \leq d(\vec{x}, \vec{y}_j), \qquad \text{con} \qquad j \neq i\} \tag{4.8}$$

es decir,

$$\mathrm{q}(\vec{x}) = \vec{y}_i, \qquad \text{si y solo si} \qquad d(\vec{x}, \vec{y}_i) \leq d(\vec{x}, \vec{y}_j),\ j \neq i,\ 1 \leq j \leq L \tag{4.9}$$

2. *Regla del vector código "centroide"*, procedente de determinar el decodificador óptimo dado un codificador fijo, que se verifica cuando $\vec{y}_i$ se elige tal que minimice las distorsión promedio de la celda C_i, es decir

$$\vec{y}_i = \mathrm{cent}(C_i) = \arg\min_{\vec{y}} E[d(\vec{x}, \vec{y}) | \vec{x} \in C_i) \tag{4.10}$$

Al vector $\vec{y}_i$ que cumple esta condición, se llamara *centroide.*

La demostración de ambas condiciones se puede realizar aplicando un procedimiento clásico de cálculo diferencial sobre la expresión (4.7), o bien utilizando argumentaciones más intuitivas como las proporcionadas en [Gersho92].

En general, las dos condiciones citadas no son suficientes para garantizar la optimización del cuantizador en su conjunto. No obstante, en ciertos casos (por ejemplo exigiendo que $\log f_X(\vec{x})$ sea concava [Gersho92]) se demuestra la suficiencia de ambas condiciones.

Para la construcción del diccionario se puede usar la generalización del algoritmo iterativo, localmente óptimo, propuesto inicialmente por Lloyd y publicado posteriormente en el año 1982 [Lloyd82], que en la literatura de *reconocimiento de formas* es conocido por *K-means.* Linde, Buzo y Gray propusieron la utilización de dicho

algoritmo (conocido también por LBG) incluso cuando la medida de la distorsión no sea una métrica [Linde80],[Buzo80]. Básicamente el algoritmo consiste en dado un diccionario inicial, aplicar iterativamente las reglas del vecino más próximo y del centroide. Por construcción, el algoritmo es descendente, en el sentido que siempre la distorsión promedio global (4.7) en cada iteración será menor o igual que en la iteración anterior, teniéndose garantizado un mínimo local tras realizar un suficiente número de iteraciones.

El carácter local hace alusión a que el mínimo conseguido depende del punto de partida o diccionario inicial considerado. En las mismas referencias, se propone la iniciación por división o *splitting*, donde se parte de un único vector código inicial (que cumple la condición de centroide para toda la secuencia de entrenamiento) que se va dividiendo progresivamente en cada paso, y por cada paso se realiza un algoritmo LBG, realizando tantos pasos como sean necesarios hasta alcanzar el número de celdas deseado.

Como alternativa al anterior y en un intento de soslayar el carácter local del mínimo se han propuesto otros algoritmos que pretenden salir de los mínimos locales alcanzados. La idea básica consiste en romper el carácter descendente del algoritmo LBG, introduciendo cierta aleatoriedad en la evolución del sistema. Se trata pues de admitir incondicionalmente perturbaciones del estado del sistema siempre que impliquen una distorsión global promedio menor y admitir probabilísticamente aquellas iteraciones que temporalmente impliquen un distorsión global promedio mayor, pero que a la postre consigan un codificador-decodificador mejor. Para estas iteraciones la probabilidad de aceptación es inversamente proporcional a lo próximo que esté el sistema del mínimo global.

Dentro de esta filosofía cabe citar los procedimientos de *"Simulated Annealing"* [Rose90] y Relajación Estocástica [Zeger92] propuestos recientemente.

4.4.2 Medida de Distancia Utilizada

En el diseño del diccionario y en la codificación, se necesita la definición de un criterio objetivo de fidelidad o medida de la distorsión. En el capítulo 3 se ha hecho un resumen de las posibles evaluaciones objetivas de calidad. Dichas medidas podrían utilizarse tanto en la fase de diseño como en la fase de codificación de un VQ. Ahora bien por cuestiones de complejidad, se descarta la distorsión espectral (apartado (3.6)) y como alternativas quedarían la razón de semejanza (o sus variantes) y la distancia cepstral.

No obstante, aun cuando dichas medidas de distancia entre las envolventes espectrales son adecuadas y coherentes con el análisis LPC, se utilizará una distancia distinta a las anteriores, para la que será fácil introducir un criterio perceptual en su definición.

Tras varias pruebas experimentales informales, se ha optado por la utilización de una distancia donde el pesado espectral introducido mejore la calidad subjetiva resultante. En particular, usando el espacio de representación de los parámetros LSP, la distancia IHM *"Inverse Harmonic Mean"* [Laroia91], definida como una distancia MSE ("mean squared error") pesada, se ha mostrado adecuada para su uso tanto en la fase de diseño como en la fase de funcionamiento del VQ. La definición del vector de pesos $\vec{P} = \{P_i\}$ con $i = 1, \cdots, p$, viene dada por la expresión

$$P_i = \frac{1}{\omega_i - \omega_{i-1}} + \frac{1}{\omega_{i+1} - \omega_i} \tag{4.11}$$

donde P_i es el peso del $i-$ésimo parámetro LSP ω_i.

La justificación del pesado realizado reside en el hecho de que la aparición de las frecuencias formantes [Soong88] es inversamente proporcional a la diferencia entre los LSPs adyacentes. Además es conocido que el oído es más sensible en las zonas formantes que en los valles espectrales; por eso, a la hora de medir la similitud entre dos espectros se pesa más estas zonas, cuando las diferencias entre los LSPs adyacentes sean menores, expresión (4.11).

bits por vector	bit-rate espectral (bps)	*DE* (dB)		*DC*		*DISNG*	
		IHM	MSE	IHM	MSE	IHM	MSE
12	533.3	2.69	2.78	2.39	2.46	0.50	0.53
11	488.8	2.85	2.94	2.56	2.62	0.54	0.57
10	444.4	3.02	3.09	2.73	2.78	0.58	0.62
9	400.0	3.20	3.27	2.93	2.98	0.64	0.68
8	355.5	3.39	3.47	3.12	3.18	0.71	0.75
7	311.1	3.64	3.73	3.38	3.46	0.79	0.83
6	266.6	3.89	4.00	3.63	3.73	0.89	0.93
5	222.2	4.21	4.29	3.96	4.04	1.04	1.07

Tabla 4.3: Distorsiones Promedio para un VQ con Distorsiones IHM y MSE.

4.4.3 Resultados Experimentales

Con los corpus de entrenamiento y test explicados en el apartado (3.5) y las condiciones de análisis mostradas en el apartado (2.4), utilizando las distancias IHM y MSE, con un procedimiento localmente óptimo LBG con iniciación por división, se han obtenido los resultados experimentales mostrados en la tabla (4.3). Se muestran resultados para distintos número de bits por vector (ó tamaño del diccionario), estimándose las siguientes medidas de distorsión:

- *DE*: Distorsión Espectral promedio. Expresión (3.6).
- *DC*: Distancia Cepstral promedio. Expresión (3.17).
- *DISNG*: Distorsión Itakura-Saito normalizada en ganancia promedio. Expresión (3.18).

Dichos resultados serán tomados como un punto de partida para el resto del trabajo, ya que pueden considerarse como referencia de un compromiso Rate-Distorsión a mejorar.

En la tabla (4.3), comparando los resultados para la distancia IHM frente a la MSE sin pesar, se observa que la primera siempre es favorable. Este resultado

bits por vector	bit-rate espectral (bps)	Mejora *DE* (dB)	Mejora *DC*	Mejora *DISNG*
10	444.4	0.007	0.004	0.002

Tabla 4.4: Mejoras Obtenidas Usando un Procedimiento de Relajación Estocástica.

"objetivo" ha sido también corroborado "subjetivamente", ya que tras varias pruebas informales, siempre se prefirió la utilización del pesado IHM frente a la utilización de una distancia MSE sin pesar.

Además de los resultados anteriores, para las secuencias de entrenamiento y test consideradas se ha introducido en el algoritmo LBG un paso adicional, que consiste en una aproximación a la relajación estocástica consistente en que a los vectores obtenidos por aplicación de la condición de centroide, se suma una componente de ruido uniformemente distribuido, cuya varianza se hace disminuir dependiendo del número de iteración actual m, según la siguiente expresión,

$$\sigma^2_{xact} = \sigma^2_{xant}\left(1 - \frac{m}{I}\right)^p \tag{4.12}$$

donde I es el número de iteraciones, m es la iteración actual y σ^2_{xact} , σ^2_{xant} son las varianzas actual y anterior respectivamente.

Para el caso particular de considerar $p = 3$ e $I = 20$ la mejora obtenida para la distorsión global en la fase de diseño del diccionario es 0.114 dB. Lo cuál se traduce en una mejora casi despreciable frente al algoritmo localmente óptimo LBG, como se indica en la tabla (4.4), cuando se evalúa el mismo test que el usado para la obtención de los resultados de la tabla (4.3).

Luego, aunque localmente óptimo, y debido a las mejoras no significativas introducidas a costa de incrementar la complejidad, se considerará como punto de partida o referencia los resultados mostrados en la tabla (4.3), abandonando la aproximación estocástica para el resto del trabajo.

4.5 VQ de coste reducido

Uno de los problemas evidentes de la VQ son las necesidades tanto en coste computacional como de almacenamiento. Para un diccionario de búsqueda exhaustiva los costes crecen exponencialmente con el número de bits, llegando a ser no realizables si el número de bits aumenta considerablemente. En particular para codificar la voz, rara vez se usan diccionarios de más de 12 bits.

Para aumentar el número de bits del diccionario es necesario introducir procedimientos de búsqueda rápida que reduzcan el coste computacional asociado para el diccionario resultante. Estos procedimientos no son más que formas de introducir un cierto grado de jerarquización en el diccionario, que permita reducir el número de operaciones o búsquedas. En particular son dignos de mención los VQ con estructura en árbol *("Tree-Search VQ")* (TSVQ) y VQ multi-etapa *("Multi-Stage VQ")* (MSVQ). Ambas aproximaciones son similares en el sentido de que comparten un procedimiento de aproximaciones sucesivas.

4.5.1 VQ con estructura en árbol (TSVQ)

Como su nombre indica consiste en organizar el conjunto de centroides, de forma que el diccionario resultante tenga una estructura jerárquica, donde para cada centroide *"padre"*, en cada uno de los niveles menos el último, se conozcan cuales son los centroides *"hijos"*.

Para codificar un vector dado, se parte del centroide raíz (primer nivel del árbol) y se incrementa el nivel de profundidad hasta llegar al último nivel; seleccionando de entre todos los hijos del nodo actual, aquél que verifique la condición de *vecino más próximo* al vector de entrada. Una vez determinado el hijo del penúltimo nivel, se transmitirá la palabra código que caracterize o bien el camino seguido o bien el vector correspondiente al último hijo encontrado.

De esta manera se puede conseguir un procedimiento rápido de búsqueda, ya que se reduce el número de distorsiones a calcular. En particular, para el caso de

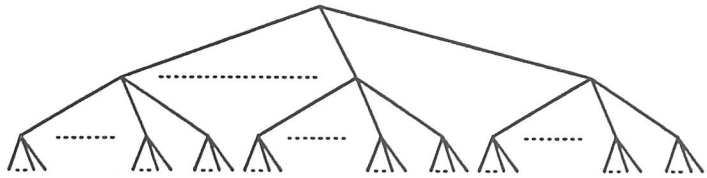

Figura 4.1: Estructura Básica de un Codificador TSVQ.

que el número de *"hijos"* de cada *"padre"* sea constante e igual a 2 (árbol binario), los L cálculos de distorsión para el diccionario de búsqueda exhaustiva se reducen a $2\log_2 L$. El precio pagado es un incremento en las necesidades de almacenamiento en el codificador, pues de los L centroides almacenados, ahora se necesitan almacenar $2*(L-2)$. Además del incremento en necesidades de memoria, este procedimiento puede introducir un empeoramiento frente al diccionario de búsqueda exhaustiva.

No siempre el árbol ha de ser binario, y no siempre el número de *hijos* para cada *padre* ha de ser constante. De hecho en [Makh85] se propone un procedimiento para construir un árbol binario donde no todos los nodos para cada nivel son expandidos (árbol *no uniforme*), que obtiene mejores resultados que un árbol uniforme.

En [Chou89] se propone un algoritmo óptimo para, dado un árbol inicial uniforme, ir *podando* recursivamente ramas, tal que el árbol no uniforme resultante siempre implique el mejor compromiso entre el *bit-rate* conseguido (en este caso variable) y la distorsión. Éste es un ejemplo de un problema convexo, y para su resolución en la referencia anterior se proporciona un algoritmo basado en multiplicadores de Lagrange, que puede ser generalizable para resolver problemas donde haya que buscar el mejor compromiso entre dos funcionales tal que uno de ellos sea monótonamente creciente y el otro monótonamente decreciente.

Finalmente en [Lopez90a], dado un diccionario inicial, se propone un algoritmo para construir eficazmente de forma iterativa un árbol a partir del diccionario de búsqueda exhaustiva de partida. Con el citado procedimiento, se puede obtener una menor distorsión menor que para el caso de la construcción del árbol mediante un procedimiento clásico LBG.

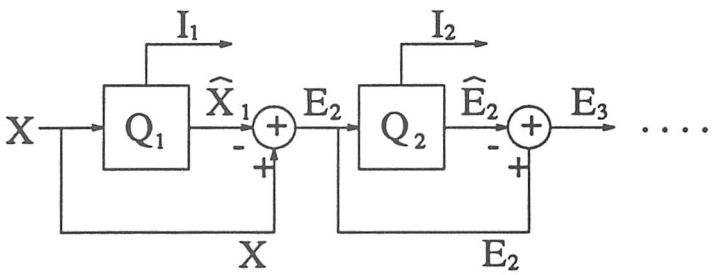

Figura 4.2: Codificador MSVQ.

4.5.2 VQ multi-etapa (MSVQ)

Este tipo de procedimientos, además de reducir el coste computacional, reducen las necesidades de memoria. Al igual que los TSVQs, codifican basándose en un procedimiento de aproximaciones sucesivas, pero ahora, en cada paso se codifica el error de cuantización (o residuo) introducido en el paso anterior, figura (4.2). También son llamados cuantizadores en cascada, ya que la decodificación se obtiene como suma de los vectores código decodificados para cada etapa, figura (4.3). Para el diseño de un codificador MSVQ, tradicionalmente se aplica un algoritmo LBG en cada etapa para la secuencia de vectores de error de la etapa anterior.

Para un MSVQ el coste computacional y las necesidades de memoria serán $\sum_{i=1}^{N_e} L_i$, en vez de los $\prod_{i=1}^{N_e} L_i$ necesarios para el caso de una búsqueda exhaustiva y una sola etapa, siendo N_e el número de etapas consideradas y L_i el tamaño del diccionario para la etapa i.

El precio pagado por esta reducción es evidentemente la obtención de unas prestaciones del sistema resultante peores comparadas con un diccionario de búsqueda exhaustiva. La razón reside en que tal como se ha indicado, en el procedimiento de diseño se está suponiendo que todos los errores cometidos en una etapa tienen una distribución común, independientemente del centroide elegido en la etapa anterior, lo cual no es cierto; este hecho es responsable de la introducción de distorsión adicional frente a un diccionario de búsqueda exhaustiva. Para resolver este problema, en [Bar93] se proponen las condiciones que debe cumplir y cómo diseñar un MSVQ

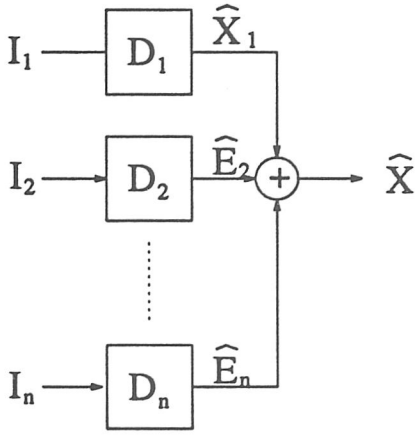

Figura 4.3: Decodificador MSVQ.

localmente óptimo, si bien, el número de operaciones necesarias excede al número necesario para un diccionario de búsqueda exhaustiva, por lo que normalmente se usan MSVQ subóptimos.

4.5.3 Códigos Producto. VQ por división (Split-VQ)

Otra aproximación para hacer *tratable* la utilización de la VQ cuando el número de bits necesarios es alto es simplemente descomponer el vector de entrada en subvectores de menor dimensión, tal que sean *tratables* individualmente.

Para que la descomposición realizada no introduzca una perdida o desaprovechamiento de las dependencias estadísticas presentes en el vector original, ha de realizarse de forma que los subvectores resultantes sean estadísticamente independientes unos de otros. Una vez realizada la descomposición cada subvector será codificado, y en el receptor tras la decodificación de cada uno de los subvectores ha de recomponerse la versión decodificada del vector original. Un VQ producto es, por tanto, aquél que para cada vector de entrada genera $I_1, I_2, \cdots I_n$ índices, correspondientes a los n subvectores del vector original. El diccionario producto resultante será el formado por el producto cartesiano de los subdiccionarios componentes considerados. Por tanto el diccionario final tendrá un tamaño o cardinal $\prod_{i=1}^{N} N_j$ y la

bits por subvector	bit-rate espectral (bps)	*DE* (dB)	*DC*	*DISNG*
12	1066.6	1.60	1.38	0.32
11	977.7	1.78	1.54	0.35
10	888.8	1.98	1.73	0.38
9	799.9	2.20	1.93	0.43
8	711.1	2.45	2.17	0.48
7	622.2	2.72	2.42	0.54
6	533.3	3.04	2.74	0.63
5	444.4	3.36	3.06	0.74
4	355.5	3.83	3.53	0.93

Tabla 4.5: Distorsiones Promedio para un Split-VQ.

complejidad será proporcional a $\sum_{i=1}^{N} N_j$.

Evidentemente la forma más directa y sencilla de realizar un VQ producto es simplemente dividir el vector original en dos o más subvectores, codificacando cada subvector independientemente con un diccionario diferente para cada uno de las divisiones consideradas. Esta simplificación subóptima para reducir el coste computacional y de memoria es llamada "*split-VQ*" en [Pali91] y es aquí mencionada porque será utilizada posteriormente en el desarrollo de este trabajo.

En la tabla (4.5) se muestran los resultados obtenidos, considerando sólo dos subvectores. El primero siempre formado por los 4 primeros coeficientes LSP y el segundo con los 6 restantes. La justificación de la división adoptada reside en que, como ya se comentó (ver tabla 2.1), los primeros coeficientes LSP tienen una sensibilidad espectral mayor. Para la construcción de los subdiccionarios se ha realizado un algoritmo LBG usando la distancia euclídea pesada IHM, con pesos dados por la expresión (4.11).

Comparando las tablas (4.3) y (4.5) puede observarse la degradación introducida al considerar dos subvectores por cada vector original y cómo la no utilización completa de las correlaciones intra-vector influye negativamente en los resultados

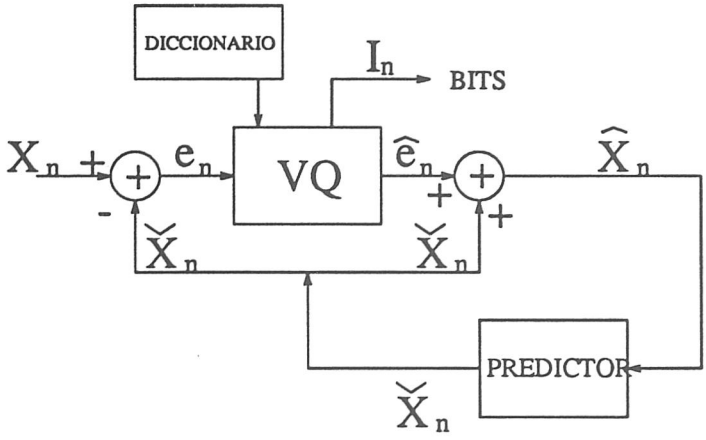

Figura 4.4: Codificador VQ Predictivo.

para el split-VQ comparado con un VQ de búsqueda exhaustiva. Por otro lado, hay que resaltar que aunque subóptimo, con split-VQ se pueden conseguir distorsiones aceptables sin necesidad de utilizar diccionarios de tamaño inmanejable.

4.6 Cuantización Vectorial Predictiva

Tras haber visto procedimientos para reducir los costes tanto computacionales como de requerimientos de memoria para un VQ, en este apartado la motivación será bien distinta, y no es otra que, dejando a un lado las cuestiones de complejidad, aprovechar la correlación todavía existente entre vectores adyacentes en el tiempo. En otras palabras, cómo aprovechar la memoria existente en la fuente de forma tal que se reduzca la velocidad de bits transmitidos por segundo.

Para ello, en este caso, en vez de cuantizar la información de entrada directamente, caracterizada por un vector de varianzas $\vec{\sigma}^2_{xi}$, se cuantizarán los errores de predicción que tendrán un $\vec{\sigma}^2_{ei}$ menor, de acuerdo con el esquema de la figura (4.4).

Ya que, para una velocidad de bits-por-segundo dada, la varianza del error de cuantización es proporcional a la varianza a la entrada del codificador, reduciendo la varianza a la entrada se conseguirá una varianza del error de cuantización menor, teniendo así un cuantizador con mejor SNR.

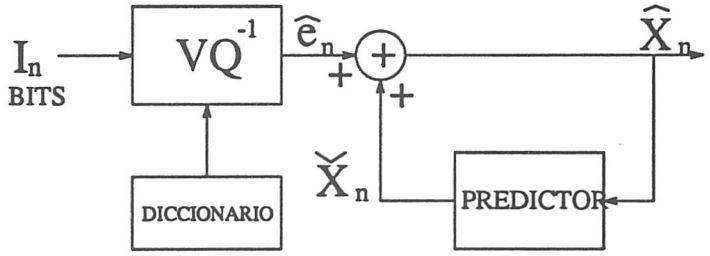

Figura 4.5: Decodificador VQ Predictivo.

Siguiendo el esquema de la figura (4.4), tras realizar una estimación vectorial $\check{x}_n$ del vector de entrada[1] x_n basándose en los vectores de reconstrucción $\{\hat{x}_k = \hat{e}_k + \check{x}_k, k < n\}$, se codificará el vector diferencia $e_n = x_n - \check{x}_n$, dando lugar a su versión cuantizada $\hat{e}_n = q(e_n)$, codificada por el índice I_n. El esquema mencionado no es más que una generalización para el caso vectorial de la DPCM *("Differential Pulse Code Modulation")*, para el que es fácilmente demostrable que

$$x_n - \hat{x}_n = e_n - \hat{e}_n \tag{4.13}$$

Es decir, el error de cuantización obtenido entre el vector de entrada y el reconstruido es igual al error de cuantización del error de predicción.

Al aplicar un VQ predictivo, cada parámetro del vector no es estimado solo en función del correspondiente parámetro en vectores anteriores sino que también es función de todos los parámetros de los anteriores vectores.

Para la evaluación de un cuantizador VQ predictivo se define la SNR, como

$$SNR_{TOTAL} = 10\log_{10}\frac{\sigma_{xi}^2}{\sigma_{ri}^2} = 10\log_{10}\frac{E[\|x_n\|^2]}{E[\|x_n - \hat{x}_n\|^2]} \tag{4.14}$$

donde σ_{ri}^2 es el vector de varianzas del error de reproducción y donde se ha supuesto que el proceso vectorial de entrada x_n es de media cero.

Teniendo en cuenta la expresión (4.13), la SNR_{TOTAL} se puede descomponer

[1] En este apartado, se notará por sencillez al vector $\vec{x}_n$ por x_n, y en general para todos los vectores, se omitirá la flecha utilizada hasta ahora.

en la contribución de dos términos, una relación señal ruido de predicción $(SNR)_P$ más un término correspondiente a la relación señal ruido de cuantización $(SNR)_Q$

$$SNR_{TOTAL} = SNR_P + SNR_Q \tag{4.15}$$

dadas por

$$SNR_P = 10\log_{10}\frac{E[\|x_n\|^2]}{E[\|e_n\|^2]} \tag{4.16}$$

$$SNR_Q = 10\log_{10}\frac{E[\|x_n\|^2]}{E[\|e_n - \hat{e}_n\|^2]} \tag{4.17}$$

Es de esperar que, si bien la relación señal ruido para un VQ codificando los errores de predicción e_n es menor que la que se obtendría al codificar los vectores de entrada x_n (debido a la menor correlación intra-trama en los e_n), la SNR_{TOTAL} del VQ predictivo sea mayor debido a que el término de la ganancia de predicción no solo corrige la disminución anterior, sino que la supera. Esta mejora (expresada en términos de la SNR_P) será tanto mayor cuanto mayor memoria exhiba la fuente.

4.6.1 Diseño del Predictor Lineal Vectorial

Dada una secuencia de vectores $\{x_n, n = -\infty, \cdots, \infty\}$, donde cada x_n es un vector m-dimensional de media cero $(E[x_n] = 0)$, el predictor óptimo es aquél para el que se cumple que la predicción $\hat{x}_n$ verifica

$$\hat{x}_n = E[x_n | x_{n-1}, x_{n-2}, x_{n-3}, \cdots] \tag{4.18}$$

Obviamente la ecuación (4.18) es impráctica fundamentalmente debido a que no se dispone de la función densidad de probabilidad condicional conjunta.

En general un predictor lineal vectorial óptimo será aquel que minimice la siguiente expresión

$$D = \sum_n E(\|e_n\|^2) = \sum_n E(\|x_n - \hat{x}_n\|^2) = \sum_n E(\|x_n + \sum_{j=1}^{m} \mathrm{A_j} x_{n-j}\|^2) \tag{4.19}$$

Donde $\mathbf{A_j}$ denota a las matrices de predicción, con dimensión $p \times p$, que caracterizan al predictor y que dependerán de la distribución estadística de la fuente. Por cuestiones de complejidad, aquí sólo se considera el caso simplificado de dependencias lineales de primer orden, con lo que en la expresión anterior la sumatoria en j queda reducida a un solo término.

La minimización de la expresión (4.19) se verifica cuando

$$\mathbf{A} = \mathbf{R}_{01}\mathbf{R}_{11}^{-1} \tag{4.20}$$

donde

$$\mathbf{R}_{ij} = E[x_{n-i}x_{n-j}^T] \approx \frac{1}{N}\sum_{n=1}^{N} x_{n-i}x_{n-j}^T \tag{4.21}$$

donde, como siempre E denota el operador de valor esperado y N es el número de vectores en la secuencia.

4.6.2 Diseño de un VQ Predictivo

Para el diseño de un cuantizador vectorial predictivo es necesario tanto determinar el predictor como construir el diccionario VQ que cuantiza los errores de predicción correspondientes. En este apartado se supone la existencia del predictor y se proporcionan una serie de procedimientos, que aun siendo subóptimos, proporcionan una manera sencilla de completar el diseño del cuantizador.

Diseño en Bucle Abierto

Éste es el procedimiento de diseño más sencillo y a la vez menos apropiado, ya que en él se adoptan las simplificaciones más drásticas. La idea es considerar la secuencia de entrenamiento como una observación empírica de la estadística de la fuente y obtener a partir de ésta una secuencia de vectores errores de predicción o residuos ideales, dados por

$$r_n = x_n - P(x_{n-1}, x_{n-2}, \cdots, x_{n-m}) \tag{4.22}$$

donde se ha notado con $P(x_{n-1}, x_{n-2}, \cdots, x_{n-m})$ al vector predicho por el predictor lineal vectorial de orden m. Una vez obtenida la secuencia de errores de predicción $\{r_n, n = 1, \cdots, L\}$, se diseña el cuantizador VQ, por ejemplo con un algoritmo LBG. A continuación se forma el sistema de la figura (4.4) que opera sobre $\hat{x}_n$, para obtener la predicción de la siguiente entrada, cerrando así el bucle.

Es obvio que este predictor no es óptimo ya que no fué diseñado para los vectores $\hat{x}_n$. Además el VQ está operando sobre los vectores diferencia e_n en bucle cerrado que no necesariamente tiene la misma distribución estadística que la secuencia de entrenamiento $\{r_n\}$.

Diseño en Bucle Semi-Cerrado

En este caso, partiendo de un predictor y un diccionario ya diseñado (que puede ser el obtenido en bucle abierto), se obtendrá una secuencia fija de residuos $\{r_n\}$, codificando la secuencia de entrada $\{x_n\}$. A partir de la secuencia $\{r_n\}$ obtenida en bucle cerrado se rediseña el diccionario en bucle abierto para la secuencia mencionada.

Diseño en Bucle Cerrado

Al igual que en las aproximaciones anteriores, se parte de la existencia de un predictor que permanecerá inalterable a lo largo de todas las iteraciones de este procedimiento. El diseño consiste en supuesto un VQ inicial (que puede ser el obtenido en bucle semi-cerrado), codificar la secuencia de entrenamiento, usando el esquema de la figura (4.4).

Ahora bien, para cada iteración, los vectores residuo e_n que fueron asignados a un vector código en particular (usando la condición de vecino más próximo) se agrupan; y posteriormente, el centroide es actualizado consecuentemente con todos los vectores e_n asignados a él [Cuper85].

Aquí, al contrario que en los procedimientos anteriores, para cada iteración la secuencia de entrenamiento es distinta, ya que el diccionario es actualizado en cada paso. Este hecho hace que la distorsión global promedio, ecuación (4.7), no sea

Diseño	*DE(dB)*	*DC*	*DISNG*
Bucle Abierto	2.62	2.29	0.44
Bucle Semi Cerrado	2.54	2.20	0.41
Bucle Cerrado	2.51	2.18	0.41

Tabla 4.6: Distorsiones Promedio para un VQ-10 bits Predictivo.

necesariamente monótonamente decreciente. No obstante, los resultados obtenidos confirman un mejor comportamiento del VQ predictivo final para el diseño en bucle cerrado que para los anteriores, como se muestra en la tabla (4.6). Para la obtención de dichos resultados, se usó la misma secuencias de entrenamiento y test que en experimentos anteriores, y el diccionario de errores de predicción se diseño con 10-bits.

Comparando las tablas (4.6) y (4.3) puede observarse la mejora introducida para el caso de un diccionario de 10 bits cuando se codifican los errores de predicción en vez de los vectores de coeficientes directamente. Por tanto, queda confirmado experimentalmente que para este caso la ganancia de predicción es mayor que la disminución en la SNR al cuantizar los errores de predicción. Además, como conclusión puede observarse que si bien para este caso los diseños en bucle semi-cerrado y cerrado mejoran al diseño en bucle abierto, las mejoras introducidas no son muy relevantes, por lo que a partir de ahora, todos los diseños que se presenten en este apartado se harán en bucle abierto, teniendo siempre presente que es de esperar una mejora poco significativa si se usaran aproximaciones en bucle semi-cerrado o cerrado.

En particular, para ver el comportamiento del esquema de la figuras (4.4) y (4.5), cuando se reduce en número de bits para la información espectral reduciendo en tamaño del diccionario, en la gráfica (4.6) se representan en el eje de abcisas la razón de bits por segundo y en el eje de ordenadas la distorsión espectral pormedio (DE). La curva etiquetada con *VQ (IHM)* hace referencia a un VQ sin predicción usando la distancia euclídea pesada IHM, mientras que la etiqueta *VQ-Predictiva*

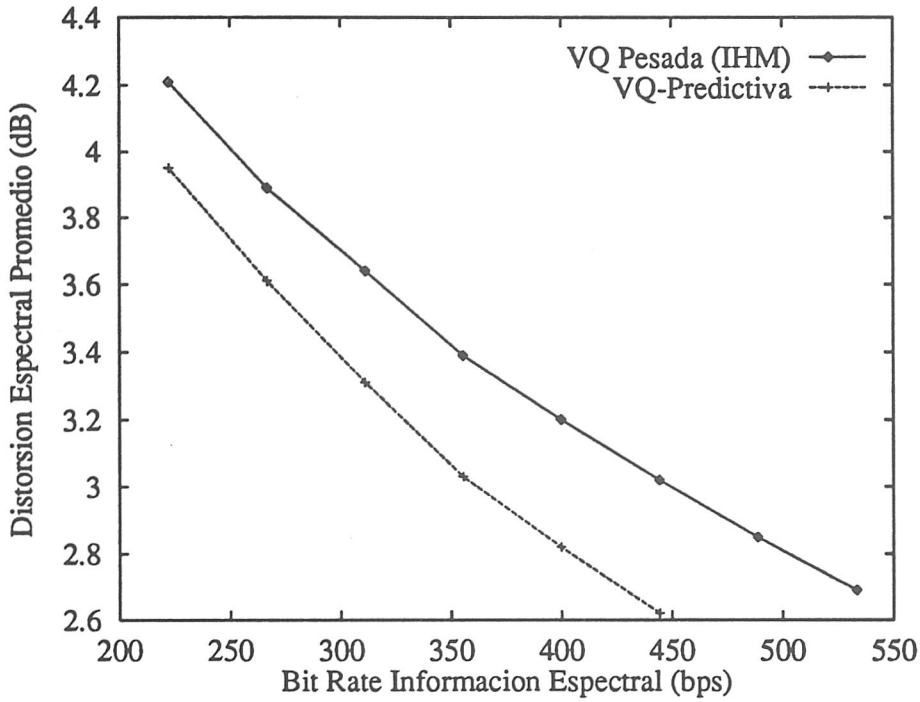

Figura 4.6: VQ con IHM Comparada con la VQ-Predictiva en Bucle Abierto.

hace referencia al esquema de la figura (4.4), habiéndose diseñado el diccionario en bucle abierto.

De los experimentos llevados a cabo, se concluye que:

- Como se observa en la gráfica (4.6), las diferencias son tanto mayores cuanto mayor es el tamaño del diccionario de predicción. De lo que se concluye que cuanto mejor es la cuantización mayor es la ganancia de predicción. Esto hace que la utilización de estos esquemas tenga más sentido en zonas de **mayores** ***bit-rates***, donde la cuantización por supuesto es mejor.

- Nótese que para mejorar el diseño, se podría actualizar el predictor e **iterativa**mente repetir los procedimientos: dado un predictor fijo rediseñar el diccionario y a continuación, fijado el diccionario actualizar el predictor en consecuencia. Esta posible refinamiento se ha llevado a cabo, pero dado lo poco significativo de las mejoras introducidas no se muestran los resultados experimentales

obtenidos.

- Por último, todos los diseños propuestos no son óptimos ya que en el diseño se está suponiendo que el vector de reproducción es perfecto. En [Chang86] se proponen diseños para los VQ predictivos basados en técnicas de gradiente descendente, para los que no se supone un vector de reproducción perfecto, obteniéndose mejores resultados. Ahora bien, la complejidad introducida en el diseño no justifica las mejoras relativamente pequeñas comparadas con los diseños subóptimos presentados.

4.6.3 VQ Predictiva Adaptable

Aunque los resultados obtenidos en el apartado anterior para un VQ predictivo son aceptables, ya que suponen una mejor calidad sin incremento en la velocidad de transmisión, hay que citar que una de sus limitaciones reside en el hecho de que sólo se usa un predictor fijo.

Es claro que toda la variabilidad de la fuente no puede ser utilizada eficientemente usando sólo un predictor, y lo que es más, es de esperar mejores resultados cuando el predictor y el diccionario de cuantización sean adaptables.

En [Fono89] se realiza una generalización de la bien conocida ADPCM ("Adaptive Differential Pulse Code Modulation") al caso vectorial, para cuando se opera en el dominio del tiempo. Hay que citar que dada la linea de investigación seguida en esta tesis, aquí tiene sólo sentido un esquema en el que no sea necesario la transmisión de información lateral (estimación "backward"), ya que de otra forma habría que caracterizar el predictor por cada vector codificado, necesitando información lateral adicional.

Ahora bien, aun usando predicción backward, dado que el objetivo es trabajar en la zona de muy baja velocidad de transmisión, no se puede obtener una buena cuantización (especialmente para diccionarios por debajo de 10-bits), lo que influye determinantemente en la predicción realizada. En definitiva, para la región de in-

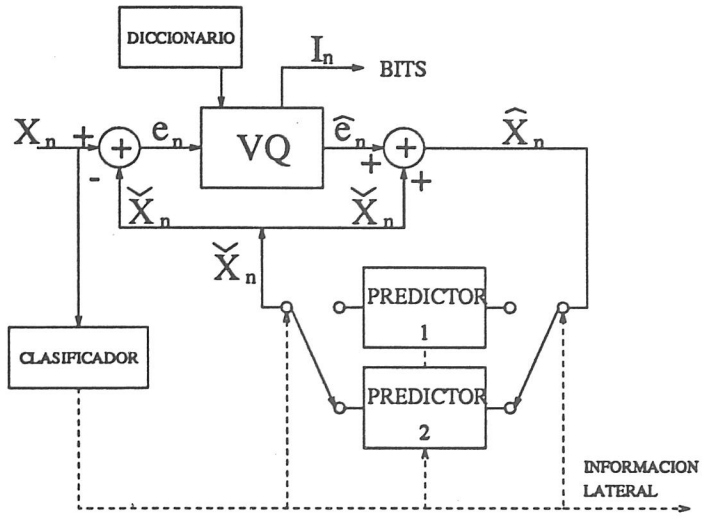

Figura 4.7: Codificador VQ Predictivo Adaptable por Conmutación.

terés, no se puede hacer una buena predicción si la cuantización realizada es pobre. Por lo tanto, al reducir drásticamente el número de bits en el diccionario no tiene sentido refinar la predicción realizada (haciéndola adaptable) porque, como ya se ha justificado, la cuantización ya es lo suficientemente inadecuada de por si.

No obstante, aquí se presentarán resultados para la utilización de una adaptación "forward" más grosera, propuesta en [Cuper85] y posteriormente utilizada en [Yong88], para cuando se opera sobre vectores de parámetros LPC.

La idea básica en esta aproximación es considerar el esquema de la figura (4.7), un VQ predictivo adaptable por conmutación, donde el objetivo es adaptarse dinámicamente a la variabilidad de la señal de entrada, manteniendo una complejidad razonable, usando más de un predictor estático.

En la fase de diseño una cuestión a resolver es cómo realizar la clasificación o la elección del predictor en bucle abierto. Hay que tener en cuenta que los predictores se caracterizan por una matriz de predicción **A**. La matriz **A** se obtuvo en función de matrices de correlación (ver expresión (4.20)). Esto sugiere hacer una clasificación que preserve ciertas propiedades de correlación para los vectores asignados a cada una de las clases. Por lo tanto aquí, habiendo calculado el vector de correlación

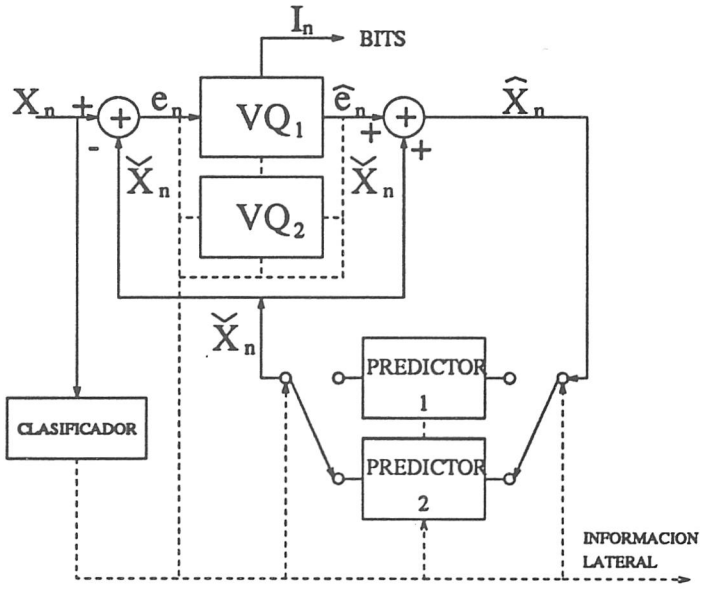

Figura 4.8: Codificador VQ Predictivo Conmutable con Cuantización Conmutable.

instantánea para el vector de entrada, se utilizará el criterio de estimar la media de sus componentes y clasificar de acuerdo con este valor según umbrales determinados experimentalmente, [Yong88].

En la fase de codificación, para el clasificador se usa un criterio "forward" exhaustivo, que no es más que elegir el predictor que minimice el error de predicción e_n, aunque podría adoptarse un criterio de selección "backward", basandonos en los vectores de reproducción $\hat{x}_n$. El precio pagado es la necesidad de información lateral adicional para caracterizar el predictor elegido, pero la ventaja es que la elección realizada no depende de la cuantización, que en nuestro caso introduce por sí sola bastante distorsión.

Además del procedimiento anterior, se introduce una modificación evidente al esquema de la figura (4.7), que no es más que cuando el número de predictores es mayor que uno y dado que se adopta un esquema "forward", hacer que el diccionario sea también conmutable, dependiendo del predictor elegido, figura (4.8), ya que sin necesidad de aumentar la información lateral transmitida, se estará realizando una mejor cuantización, a costa de un aumento de las necesidades de memoria.

Diseño	número de diccionarios	número de predictores	número total de bits	*DE(dB)*	*DC*	*DISNG*
Bucle Abierto	1 VQ-10	1	10	2.62	2.29	0.44
Bucle Abierto	2 VQ-09	2	10	2.58	2.26	0.45
Bucle Abierto	4 VQ-08	4	10	2.62	2.30	0.46
Bucle Abierto	8 VQ-07	8	10	2.69	2.37	0.48
Bucle Abierto	1 VQ-10	2	11	2.43	2.10	0.42
Bucle Abierto	2 VQ-10	2	11	2.40	2.08	0.41
Bucle Abierto	1 VQ-10	4	12	2.35	2.01	0.41
Bucle Abierto	4 VQ-10	4	12	2.26	1.93	0.39
Bucle Abierto	1 VQ-10	8	13	2.30	1.97	0.40
Bucle Abierto	8 VQ-10	8	13	2.11	1.79	0.37

Tabla 4.7: Distorsiones Promedio para un VQ Predictivo Adaptable por Conmutación con Cuantización Conmutable.

Los resultados experimentales obtenidos se muestran en la tabla (4.7). Como ya se justificó, se presentan resultados sólo para el diseño en bucle abierto. A la vista de dicha tabla se puede concluir que

- Cuando la cuantización es pobre (es decir para diccionarios de menos de 10 bits) al aumentar el número de predictores y diccionarios manteniendo el número de bits totales constante, los resultados obtenidos son cada vez peores. Una posible justificación reside en el hecho de que una mala cuantización no puede ser corregida por la predicción, ya que ésta se hace en base a aquélla.

- Por otro lado, manteniendo igualmente el número de bits totales constante, si se aumenta el número de diccionarios (de igual tamaño) para un número de predictores fijo, evidentemente los resultados son mejores como cabría esperar aunque son cada vez mejores cuanto mayor es el número de predictores.

Para concluir, en la tabla (4.8) se presentan los resultados obtenidos siguiendo el mismo esquema de la figura (4.8), pero ahora usando diccionarios por división (Split-VQ) (apartado (4.5.3)). De los resultados mostrados en dicha tabla se puede concluir que

- Comparando con un split-VQ sin predicción, con solo aumentar un bit por

Diseño	número de diccionarios	número de predictores	número total de bits	*DE(dB)*	*DC*	*DISNG*
Bucle Abierto	2 Split-VQ-22	2	23	1.26	1.05	0.27
Bucle Abierto	2 Split-VQ-24	2	25	1.12	0.93	0.26
No Predicción	1 Split-VQ-22	0	22	1.78	1.54	0.35
No Predicción	1 Split-VQ-24	0	24	1.60	1.38	0.32

Tabla 4.8: Distorsiones Promedio para un Split-VQ Predictivo Adaptable por Conmutación con Cuantización Conmutable.

vector, usando un split-VQ predictivo adaptable por conmutación con cuantización conmutable se consiguen mejoras significativas del orden de 0.5 dB de Distorsión Espectral.

4.7 Cuantización Vectorial Generalizada

En el apartado anterior se ha discutido el aprovechamiento de las dependencias lineales entre vectores adyacentes para, con ello, realizar una codificación más eficiente. Evidentemente, las dependencias estadísticas entre vectores adyacentes no se limitan al caso lineal, sino que como es de suponer, si se desarrollan técnicas que aprovechen dichas dependencias no lineales, se obtendrán esquemas de codificación-decodificación que impliquen un mejor compromiso tasa de bits-calidad.

Para el aprovechamiento de las dependencias no lineales, que un codificador escalar óptimo no puede utilizar, en el apartado (4.4) se justificó la conveniencia de la utilización de la VQ frente a los codificadores escalares. Todo la argumentación aludida puede ser trasladable aquí para justificar la utilización de una generalización de la VQ para el aprovechamiento de las dependencias no lineales entre vectores adyacentes que un VQ predictivo no puede utilizar.

La idea básica consiste pues en considerar como unidad un grupo de vectores adyacentes en el tiempo que serán cuantizados y codificados conjuntamente, dando lugar a lo que aquí se llama la *Cuantización Vectorial Generalizada.*

Dependiendo de que el número de vectores adyacentes considerados sea fijo o variable, se hablará de cuantización matricial (MQ) ó cuantización por segmentos (SegQ). Obviamente, la primera es mucho más sencilla, en términos de complejidad, que la segunda, ya que es una generalización directa de la cuantización vectorial (VQ) al caso multidimensional y como se verá, tanto el diseño como la codificación-decodificación son realizables por generalización directa. Ahora bien, debido al carácter no estático de la señal de voz, es de esperar que una generalización dinámica (es decir, considerar unidades de longitud variable) implique un mejor compromiso para el codificador-decodificador, si bien el precio pagado para este caso (SegQ) es la utilización de una tasa de bits variable, un retardo variable y por supuesto un aumento en la complejidad tanto en la fase de diseño como en la de operación.

Para hacer uso de las dependencias no lineales entre vectores adyacentes una alternativa a considerar, frente a las anteriores, sería la utilización de un VQ junto a una máquina de estados finitos *"Finite-State Vector Quantization" (FSVQ)* [Dun85], en la que el diccionario de vectores código no es único sino que depende del estado del codificador. No obstante, aquí no es considerada, ya que en la región de interés para la linea seguida en esta tesis, la MQ se compara favorablemente con la FSVQ, tanto en términos de complejidad como de calidad, [Tsao85].

4.7.1 Cuantización Matricial (MQ)

Un cuantizador matricial MQ opera sobre vectores lo mismo que un cuantizador vectorial opera sobre muestras, es decir los símbolos de la fuente (en este caso vectores) son agrupados, cuantizados y codificados como una unidad de un espacio multivectorial.

Para el diseño del cuantizador matricial, al igual que para el caso vectorial, han de verificarse las dos condiciones suficientes: *Regla de mínima distorsión o vecino más próximo* y *Regla de la matriz código "centroide"*, que utilizando un generalización del algoritmo LBG proporcionarán un mínimo local de la distorsión global promedio

para el diccionario matricial.

Un problema asociado con la MQ es que para captar todas las posibles transiciones se necesita una secuencia de entrenamiento mucho mayor que para el caso vectorial. En los resultados presentados a continuación, se ha optado por utilizar como secuencia de entrenamiento para el diseño del diccionario MQ la misma secuencia que para el caso vectorial, pero considerando como matrices de entrada todas las resultantes de ir incrementando en uno el puntero que direcciona el primer vector de la matriz. Es decir la j-ésima matriz de la secuencia estará formada por los vectores[1] $\mathbf{X_j} = [x_j, x_{j+1}, \cdots, x_{j+L-1}]$, siendo L el número fijo de vectores que forman la matriz.

Además, para introducir un criterio subjetivo en la distancia, se usa una generalización de la distancia IHM al caso matricial, dada por

$$d(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{L} d_{IHM}(x_i, y_i) \tag{4.23}$$

donde la $d_{IHM}(x_i, y_i)$ es la distorsión IHM para el caso vectorial, con pesos descritos por la expresión (4.11).

En la gráfica (4.9) se muestran los resultados obtenidos, considerando valores de $L = \{2, 3, 4\}$, etiquetados respectivamente con $MQ \quad L = \{2, 3, 4\}$. Para un L dado, se obtienen distintos valores del *bit-rate* haciendo variar el tamaño de los diccionarios. Para comparación se incluyen los resultados obtenidos para la gráfica (4.6), etiquetados de nuevo con *VQ-Predictiva* y *VQ Pesada (IHM).*

Observando los resultados obtenidos en la gráfica (4.9) se puede concluir que

- Para bit-rates inferiores a 300 bps, con la generalización de la VQ a MQ se obtienen mejores resultados que con la VQ-predictiva en bucle abierto y por supuesto que la VQ. Esta mejora sustancial, se puede atribuir a que con la MQ se están aprovechando más eficazmente las dependencias estadísticas que

[1] Para la cuantización vectorial generalizada se notará a las unidades de cuantización usando caracteres en negrita y mayúsculas

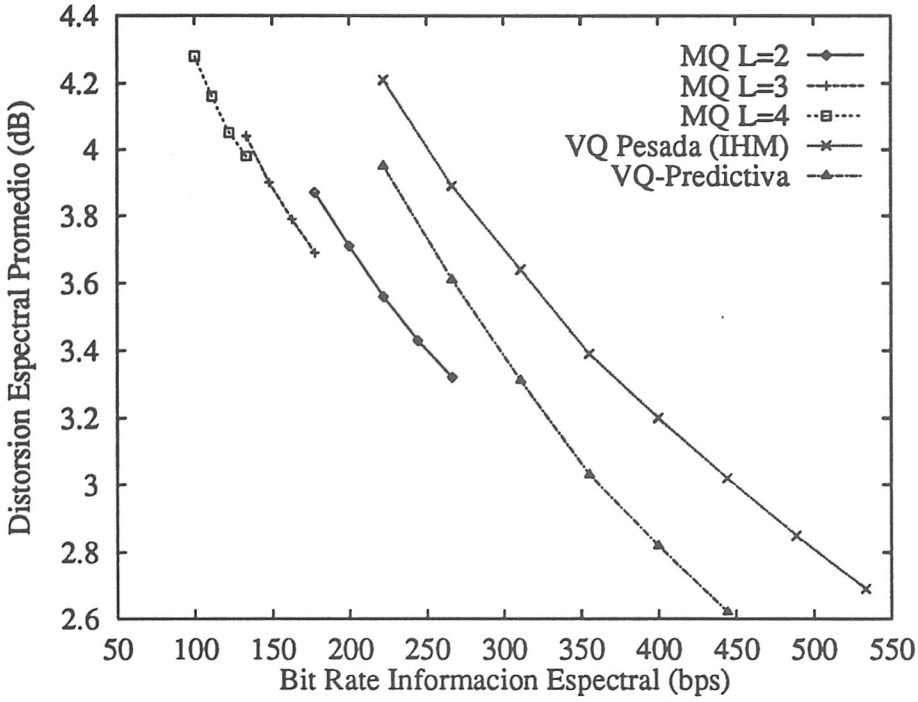

Figura 4.9: MQ con IHM comparada con la VQ-Predictiva en Bucle Abierto y VQ con IHM.

exhibe la fuente, en particular las dependencias no lineales que un predictor lineal no puede aprovechar.

- Al aumentar el tamaño de la matriz L se observa por supuesto una tendencia hacia la zona de bit-rates menores, acompañado de un incremento de la distorsión.

- Para una distorsión dada, se obtienen mejores resultados cuando L es mayor, ahora bien, la mejora en términos relativos es progresivamente menor cuanto L es mayor, lo que indica que pocas dependencias se pueden esperar entre vectores separados más que la duración de 3 tramas, o sea más de 67.5 ms.

- Para la obtención de los puntos en las tres curvas con distorsión espectral menor, se han usado diccionarios matriciales de 12 bits. Una reducción de la distorsión se podría conseguir considerando diccionarios mayores, pero en este

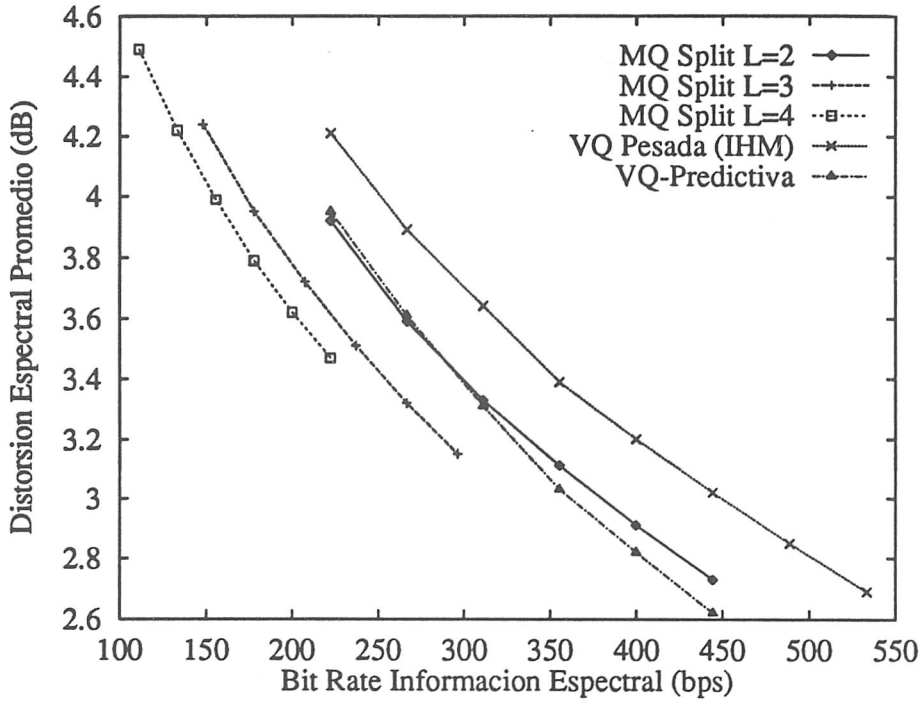

Figura 4.10: MQ-Split con IHM Comparada con la VQ-Predictiva en Bucle Abierto y VQ con IHM.

caso la complejidad computacional necesaria los haría inmanejables.

Como se ha citado en el último punto, un intento de mejorar los resultados puede venir de hacer una cuantización mejor de la matriz, es decir usar más información para caracterizarla. Evidentemente, por cuestiones de complejidad, no es posible tanto para la fase de diseño como para la fase de operación aumentar sin más el número de bits para el MQ. Como posibilidades se podrían considerar tanto una generalización de la VQ multi-etapa (MSVQ) ó una generalización matricial del Split-VQ (apartado (4.5.3)). En la gráfica (4.10) se muestran los resultados para la generalización realizada, etiquetada con *MQ-Split*, para el caso de $L = \{2, 3, 4\}$. Para la codificación de cada matriz se consideran dos diccionarios, uno para codificar la submatriz formada por los cuatro primeros parámetros de los L vectores considerados, y un segundo diccionario independiente, formado por los seis restantes parámetros de los L vectores considerados.

A la vista de los resultados de la gráfica (4.10) se concluye que

- Para el caso de $L = 2$, con un MQ-Split se obtienen similares resultados que con un cuantizador predictivo en bucle abierto para bit-rates menores de 300 bps. Ahora bien, las diferencias sustanciales que se observaban en la gráfica (4.9) entre la MQ y la VQ-predictiva se han perdido en este caso, pues aquí, con la simplificación realizada, claramente se está desaprovechando la correlación intra-trama. En definitiva, las dependencias no-lineales entre vectores adyacentes son de igual magnitud que las dependencias intra-trama.

- El comportamiento se invierte para bit-rates mayores de 300 bps. Esto se puede atribuir al problema de entrenamiento insuficiente en el diseño de MQ; que como es de esperar, se pone más de manifiesto cuando el tamaño del diccionario es mayor.

- Se observa que aumentando el tamaño L de las *split-matrices* se obtienen mejores resultados, como ya se había constatado para la MQ, indicando que las dependencias estadísticas exhibidas por la fuente van más allá de afectar sólo a vectores adyacentes. Ahora bien, igualmente que para el caso MQ, se observa una tendencia a la saturación al aumentar L, indicando que las dependencias inter-trama no van más allá de 3 o 4 tramas.

4.7.2 Cuantización en Segmentos (SegQ)

En este apartado se considera otra aproximación a la generalización de la VQ a más dimensiones. Al igual que en el esquema MQ, se toma como unidad de entrada al codificador la unión de varios vectores adyacentes, pero ahora en un número variable. La justificación de esta aproximación reside en el hecho de que, lejos de tener un comportamiento estático, como el considerado en la aproximación matricial, las zonas de homogeneidad espectral en el habla tienen una duración variable, esto justifica que se consideren como unidades a codificar elementos de duración variable (expresado en número de vectores adyacentes a considerar) que se llamarán *segmentos*.

Con la aproximación propuesta, surge un problema a resolver que no es otro que determinar los límites de los segmentos, denominado clásicamente *segmentación*. En general, el objetivo de la segmentación está intimamente relacionado con la determinación de zonas que exhiban cierta homogeneidad tanto espectral como temporal (un buen resumen de aportaciones realizadas en los últimos años puede encontrarse en [Vidal90]), ahora bien, para el caso particular de los problemas de codificación, donde la meta última es maximizar un criterio de fidelidad, la segmentación se puede reformular como

> *la determinación de los límites de los segmentos, tal que implique una minimización de la distorsión entre la secuencia de entrada y la de salida en el codificador-decodificador.*

Al realizar la segmentación, como ya se ha indicado, se debe considerar un criterio de fidelidad o medida de la distorsión, a maximizar o minimizar respectivamente. Es necesario hacer una normalización de las duraciones de los segmentos a comparar para obtener una medida significativa de la distorsión. Caben varias posibilidades que son:

- *variable_a_fijo*, que consiste en normalizar y muestrear el segmento de entrada a una longitud fija, predeterminada e igual a la duración de los segmentos-código
- *fijo_a_variable* que por el contrario consiste en normalizar o muestrear los segmentos-código a la duración del segmento de entrada.

A su vez, la normalización de los segmentos puede realizarse tanto en

- el dominio del tiempo, muestreando la secuencia a intervalos regulares de tiempo
- el dominio de la frecuencia, muestreando la trayectoria espectral a intervalos regulares, dado algún criterio de distorsión.

Para una distorsión promedio dada, en general la SegQ pretende reducir el tamaño del diccionario, comparado con el que se necesitaría con un MQ, a expensas de un aumento en la complejidad del codificador (hay un procedimiento de segmentación incluido) y por supuesto siendo necesario la transmisión de información lateral que caracterize la duración del segmento.

Diseño del Cuantizador por Segmentos

Si se tiene un procedimiento de segmentación a priori, en el sentido de que no dependa de la cuantización a realizar, y se base simplemente en caracterizar zonas de homogeneidad espectral, como el propuesto en [Segu90], el diseño del cuantizador, supuesto un procedimiento de normalización temporal para el cálculo de distancias entre los segmentos, es una generalización directa del diseño VQ. En las referencias [Rou82], [Rou83] y posteriores, se propone esta fácil generalización.

No obstante, si se desea hacer un diseño conjunto, en el sentido de decidir cual de las segmentaciones posibles implica un distorsión promedio menor, el procedimiento no es una generalización directa del VQ, ya que tanto en la fase de diseño como de operación del codificador-decodificador habrá que realizar la segmentación y la cuantización conjuntamente mediante un procedimiento típico de alineamiento temporal, como el propuesto en [Shira88], que aquí se presenta y del que se toma la notación.

Medida de Distorsión por Segmentos. Normalización Temporal

Debido a que el cálculo de distorsiones usando una normalización fijo_a_variable de los segmentos código implica un minimización en el espacio del segmento de entrada, aquí se usará dicha normalización frente a la alternativa de usar una normalización de variable_a_fijo que supondría minimizar distorsiones en el espacio de los segmentos código y no en el del segmento de entrada.

Sea $\{\mathbf{X_1}, \cdots, \mathbf{X_j}, \cdots, \mathbf{X_J}\}$ un *bloque* de J segmentos de entrada al codificador, cada uno de ellos con longitud variable l_j (expresada en número de vectores). Sea

$C = \{\mathbf{Y}_i\}_{i=1}^{N}$ el diccionario SegQ de segmentos de longitud fija L. Si las duraciones de los segmentos del bloque de entrada son desconocidas, la regla del vecino más próximo o de mínima distorsión para dicho bloque, puede formularse como

$$\hat{\mathbf{X}}_j|_{j=1}^{J} = \arg\min_{l_j} \left\{ \min_{\mathbf{Y}_{i(j)} \in C} \sum_{j=1}^{J} d(\mathbf{X}_j, \mathbf{Y}_{i(j)}) \right\} \qquad \text{con} \qquad 1 \leq j \leq J \tag{4.24}$$

que literalmente quiere decir encontrar los límites de los segmentos dentro del bloque y los segmentos-código $\mathbf{Y}_{i(j)}$ que minimizan la distorsión. En la expresión anterior las longitudes están restringidas a todo conjunto de l_j posibles tal que su suma sea igual a la duración del bloque considerado.

La distorsión en la expresión (4.24) se evalúa haciendo una normalización temporal de los vectores-código fijo_a_variable, tomando como referencia la longitud del segmento de entrada. Para el segmento t del bloque, los L vectores del segmento código son linealmente interpolados y muestreados dando lugar a un nuevo segmento-código $\check{\mathbf{Y}}$ de longitud l_t. En notación matricial

$$\check{\mathbf{Y}} = \mathbf{Y}\mathbf{H}_{l_t} \tag{4.25}$$

donde $\mathbf{H}_{l_t}$ es la matriz L por l_t de interpolación temporal. Dicha matriz sólo depende de la longitud l_t y de L, y en cada columna tiene sólo dos elementos distintos de cero $h_{i_1 j}$ y $h_{i_2 j}$ con $j = 1, 2, \cdots, l_t$. Definiendo

$$\beta = \frac{(j-1)(L-1)}{(l_t - 1)} \tag{4.26}$$

y

$$\alpha = \beta - [\beta] \tag{4.27}$$

donde $[\beta]$ denota al mayor entero que no exceda a β, se tiene que

$$h_{i_1 j} = 1 - \alpha \qquad \text{siendo} \qquad i_1 = [\beta] + 1 \tag{4.28}$$

y

$$h_{i_2 j} = \alpha \qquad \text{siendo} \qquad i_2 = [\beta] + 2 \tag{4.29}$$

Segmentación y Cuantización usando Alineamiento Temporal

Antes de introducir la recursión utilizada, se define la siguiente notación

T Número de vectores en el bloque de entrada

J Número de segmentos en el bloque de entrada

t_j Límites de los segmentos dentro del bloque

t_j^0 Límites iniciales de los segmentos dentro del bloque

t_j^* Límites óptimos de los segmentos dentro del bloque

Δ Intervalo de búsqueda de los límites de los segmentos

$\mathbf{X}(t_{j-1}, t_j)$ Segmento con límites situados en t_{j-1} y t_j

$q(\mathbf{X}, \mathbf{Y})$ Distorsión mínima obtenida al cuantizar $\mathbf{X}$ usando el diccionario C

El procedimiento recursivo para evaluar la distorsión y las posiciones óptimas de los límites de los segmentos dentro del bloque, es el resultado de aplicar un procedimiento típico de búsqueda donde dada una segmentación inicial, se calcula la distorsión acumulada de todas las posibles segmentaciones, dentro de la restricción impuesta por Δ, y de entre ellas se elige la segmentación que implique una distorsión de cuantización menor acumulada para todo el bloque. Es decir, usando la siguiente recursión

$$\sigma(t_j) = \min_{t_{j-1} \in R_{j-1}} \{\sigma(t_{j-1}) + q(\mathbf{X}(t_{j-1}, t_j), \mathbf{Y})\} \tag{4.30}$$

$$s(t_j) = \arg \min_{\substack{t_{j-1} \in R_{j-1} \\ t_j^0 - \Delta \leq t_j \leq t_j^0 + \Delta \\ j = 1, 2, \cdots, J}} \{\sigma(t_{j-1}) + q(\mathbf{X}(t_{j-1}, t_j), \mathbf{Y})\} \tag{4.31}$$

donde

$$\sigma(t_0) = 0, \qquad t_0 = t_0^* = 1, \qquad t_J = t_J^* = T$$

$$q(\mathbf{X}, \mathbf{Y}) = \min_{\mathbf{Y}_i \in C} d(\mathbf{X}, \mathbf{Y}_i)$$

$$R_{j-1} = \{t | t_{j-1}^0 - \Delta \leq t \leq \min(t_{j-1}^0 + \Delta, t_j)\}, \qquad R_0 = 1 \tag{4.32}$$

Una vez calculada $\sigma(t_J)$, las posiciones óptimas de los segmentos se obtienen usando la siguiente recursión hacia atrás

$$t_{j-1}^* = s(t_j) \qquad j = J, J-1, \cdots, 2 \tag{4.33}$$

Para la elección de la segmentación a priori t_j^0, se puede elegir cualquier procedimiento basado en determinar zonas de homogeneidad espectral, como el propuesto en [Segu90], suponiendo que una segmentación basada en determinar zonas de homogeneidad espectral es adecuada como punto de partida para la minimización de la distorsión acumulada de cuantización del bloque. Sin embargo, aquí la segmentación a priori elegida consiste en realizar una técnica de agrupamiento o *"clustering"*, tal que los vectores en el bloque se consideran pertenecientes al segmento actual, siempre que la distorsión al "centroide" del "cluster" formado por los vectores pertenecientes al segmento esté por debajo de un umbral dado.

Es de esperar que las deficiencias de la segmentación a priori realizada, sean corregidas por el procedimiento recursivo presentado, en términos de reducir la distorsión de cuantización tras la codificación. Ahora bien, el precio pagado es la introducción de un retardo T en el codificador igual a la duración del bloque considerado $\{\mathbf{X}_1, \cdots, \mathbf{X}_j, \cdots, \mathbf{X}_J\}$.

Diseño del Diccionario por Segmentos

Suponiendo que se dispone de una secuencia de entrenamiento suficientemente grande para caracterizar la distribución estadística de los segmentos para una fuente dada,

el diseño del diccionario consistirá en determinar los segmentos de dicha secuencia y los segmentos código que impliquen una distorsión de cuantización menor acumulada para toda la secuencia de entrenamiento.

El procedimiento consiste en la iteración hasta converger de los siguientes dos pasos,

- dado un diccionario de segmentos-código actualizar los límites de los segmentos de la secuencia de entrenamiento usando el procedimiento recursivo de las expresiones (4.30) y (4.31).
- dada la secuencia de entrenamiento con segmentos calculados en el paso anterior, actualizar el diccionario de segmentos-código, tal que minimice la distorsión global promedio por segmento

En el segundo paso, ó de actualización de los vectores código, se usa la regla del segmento-código más próximo, seguida de la aplicación de la regla de obtención del segmento centroide. Dichas reglas son una generalización a más dimensiones de las reglas del vecino más próximo y del vector centroide respectivamente, usadas en el diseño VQ. Se puede demostrar fácilmente que para la partición i, el segmento centroide se calcula por

$$\mathbf{Y}_i = \left(\sum_{j \in S_i} \mathbf{X}_j \mathbf{H}_{l_t}^t \right) \left(\sum_{j \in S_i} \mathbf{H}_{l_t} \mathbf{H}_{l_t}^t \right)^+ \tag{4.34}$$

donde se ha notado $\mathbf{A}^+$ a la matriz generalizada inversa[1] de $\mathbf{A}$. La obtención de la expresión anterior, se obtiene al minimizar la distorsión D_i para cada partición S_i dada por

$$D_i = \sum_{j \in S_i} d(\mathbf{X}_j, \mathbf{Y}_i). \tag{4.35}$$

Resultados Experimentales

[1] Debido a que la matriz $\mathbf{H}_{l_t} \mathbf{H}_{l_t}^t$ puede ser singular, para su inversión se usa el método de mínimos cuadrados o Moore-Penrose [Golub83]

etiqueta	T/J
SegQ-1	2.75
SegQ-2	4.05
SegQ-3	5.67

Tabla 4.9: Duraciones Medias en Vectores por Segmento para los Distintos Umbrales Considerados.

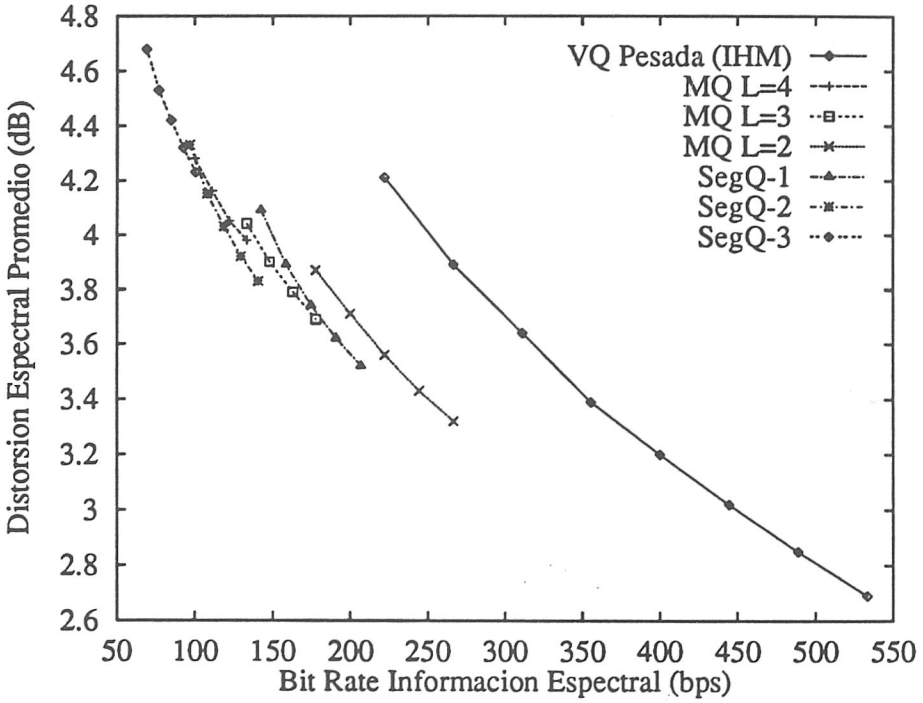

Figura 4.11: Cuantización por Segmentos con Duración Máxima 8 vectores y Retardo Infinito, MQ y VQ con IHM.

Considerando distintos umbrales para la segmentación a priori, se obtuvieron distintas duraciones promedio por segmento (T/J) para la secuencia de test. En la tabla (4.9) se muestran las duraciones consideradas para la obtención de la gráfica (4.11) expresadas en número medio de vectores por segmento, junto con la etiqueta identificativa utilizada en dicha gráfica. Para cada una de las duraciones promedio consideradas, en la gráfica (4.11) se representan los resultados obtenidos tras codificar la secuencia de test con diccionarios de tamaño $\log_2(N) = \{6, 7, 8, 9, 10\}$ bits.

Se ha utilizado un umbral de búsqueda $\delta = 3$; así mismo, la longitud fija con la que son almacenados los segmentos código del diccionario es $L = 3$ vectores. Para la segmentación a priori se ha impuesto un retardo máximo de 8 vectores por bloque (que significan menos de 200 ms para las condiciones de análisis (ver tabla 2.6)), por coherencia con las especificaciones del nuevo estándar a 2400 bps [Welch93]). Ahora bien, toda la secuencia de test se consideró como un solo bloque, con lo que en realidad el retardo es igual a la duración de dicha secuencia. En otras palabras, esta consideración limita la posible aplicación de este procedimiento, no obstante, su interés reside a la vez en fijar el límite experimental del mejor comportamiento que se puede esperar imponiendo una duración máxima de 8 vectores por segmento pero permitiendo retardos en el codificador tendiendo a infinito.

En la misma gráfica se incluyen como referencia los resultados obtenidos en anteriores apartados de este trabajo etiquetados *VQ Pesada (IHM)* y *MQ L=2,3,4*, correspondientes a un cuantizador vectorial con distancia pesada IHM (sección (4.4)) y un cuantizador matricial con $L = \{2, 3, 4\}$ (sección (4.7.1)), respectivamente.

Tras la obtención de los resultados mostrados en la gráfica (4.11) se concluye que

- Es evidente que aplicar un procedimiento como el descrito donde se busca de entre todas la posibles segmentaciones aquélla que implique un distorsión de cuantización menor (tanto en la fase de diseño como en la de codificación) puede mejorar los resultados obtenidos al considerar una segmentación fija e independiente de la cuantización a realizar (MQ).

- Aunque corregida por el procedimiento descrito, la segmentación a priori determina drásticamente las prestaciones del sistema, ya que una vez fijada, se habrá fijado el bit-rate, sin considerar ningún criterio de fidelidad.

- Como ya se ha comentado estos resultados tienen un relativo interés práctico, ya que suponen un retardo tendente a infinito (la duración de la secuencia de entrada). En la referencia [Shira88], se toman como bloques de entrada al conjunto de vectores delimitados por zonas de silencio, lo cuál aunque supone

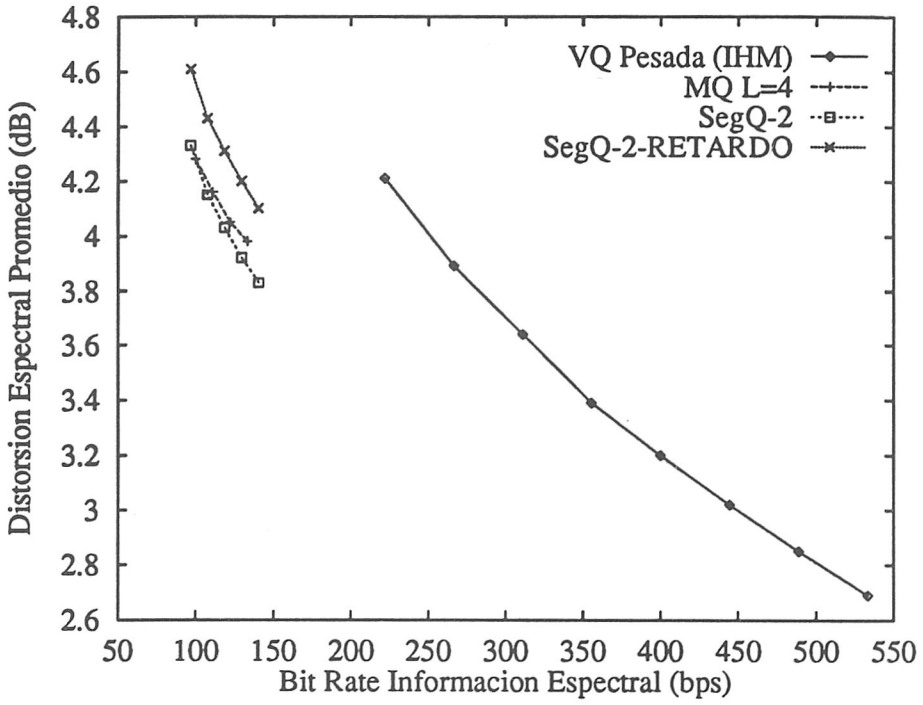

Figura 4.12: Influencia del Retardo Máximo en la Cuantización por Segmentos.

una segmentación natural, introduce un retardo que puede ser grande y, lo que es peor, variable. Por lo tanto, la aplicación de este esquema tiene sentido cuando el retardo no sea una constricción o premisa de diseño, por ejemplo en comunicaciones con un solo sentido *(simplex)* y en almacenamiento digital de voz.

- Al reducir el tamaño del diccionario se observa que el comportamiento es muy próximo o incluso peor al obtenido con un esquema MQ, lo que quiere decir que cuando la cuantización introduce una elevada distorsión no tiene sentido realizar una buena segmentación.

En el siguiente experimento se impone una limitación adicional consistente en considerar bloques de entrada con una duración máxima de 8 vectores. Luego, en este caso, tendremos un retardo dentro de las especificaciones del estándar antes citado [Welch93]. En la gráfica (4.12), se representan los resultados obtenidos para los es-

quemas etiquetados anteriormente por *VQ Pesada (IHM), MQ L=4* y *SegQ-2*, junto con el nuevo esquema, etiquetado *SegQ-2-RETARDO* donde se ha considerado un umbral para la segmentación a priori que proporcione una duración promedio, es decir un bit-rate, igual al obtenido en el esquema *SegQ-2*.

De los resultados mostrados en la gráfica (4.12) se concluye que:

- Imponer un límite máximo para la duración de los bloques igual a 8 segmentos deteriora las prestaciones del sistema, ya que ese límite ha sido fijado para acotar el retardo y no necesariamente coincide con un límite real en la secuencia de test. De hecho, se observa que esta imposición drástica implica unos resultados incluso peores que los obtenidos con un esquema MQ. Lo que indica que una segmentación con un criterio local (debido a la longitud máxima del bloque impuesta) es peor que una segmentación fija, como la utilizada en los MQ..

Limitaciones y Posibles Mejoras del SegQ

En este subapartado se comentan las limitaciones encontradas en el sistema desarrollado de cuantización por segmentos (SegQ) y se proponen procedimientos que podrían considerarse como alternativas para soslayar las limitaciones citadas, así como otros encontrados en la bibliografía.

- Como se ha explicado, para normalizar las duraciones, se realiza una conversión del segmento código (almacenado con una duración fija) a la longitud variable del segmento de entrada. Esta normalización no es más que un muestreo de la trayectoria en el espacio multidimensional a intervalos regulares de tiempo, es decir es un alineamiento temporal con interpolación. Evidentemente, no siempre esta normalización es la mejor, en el sentido de que si el segmento presenta un comportamiento poco estático (es decir, hay mucha variabilidad en cortos intervalos de tiempo) con la normalización realizada se estará suavizando dicha variabilidad. En estas situaciones es de esperar que un muestreo

"espacial" resulte mejor que el "temporal".

- Otra limitación del SegQ es que para realizar la cuantización es necesario conocer el número de segmentos a priori por cada bloque considerado y además, disponer de una segmentación inicial, que con el procedimiento utilizado, no depende de la cuantización a realizar, sino que se basa en un criterio de homogeneidad espectral que no siempre será el que minimice la distorsión de cuantización. En otras palabras, el *bit-rate* se determina basándose en un criterio que no implica maximizar el criterio de fidelidad.

  Como alternativa, se podría considerar la aplicación de técnicas clásicas en el reconocimiento de patrones como es el alineamiento temporal dinámico ("*Dinamyc Time Warping (DTW)*") [Myers81] ó [Ney84] donde no es necesario conocer el número de segmentos en el bloque de entrada. Además como ventaja adicional, utilizando un procedimiento DTW, el alineamiento temporal como su nombre indica se realiza dinámicamente (al contrario que el realizado hasta ahora consistente en un muestreo periódico de la trayectoria multidimensional) con lo que es de esperar una reducción en la distorsión de cuantización obtenida. Además, para este caso, el *bit-rate* resultante está relacionado con la maximización del criterio de fidelidad entre la entrada y la salida.

  En particular, para observar la ganancia que cabría esperar al realizar un casamiento dinámico, sin necesidad de realizar un segmentación a priori del bloque de entrada, se ha desarrollado un sistema que codifica y decodifica realizando el alineamiento temporal dinámico. Este procedimiento determina de entre todos los segmentos código y todos los alineamientos posibles, aquellos que minimicen la distorsión de cuantización (o coste) con el bloque de entrada.

  En la construcción del camino de alineamiento, dado que el coste asociado al mismo depende de su longitud, al calcular las distorsiones se utilizan los pesos descritos en la figura (4.13); la asignación de pesos realizada pretende que caminos con menor longitud sean penalizados localmente. Con la aplicación

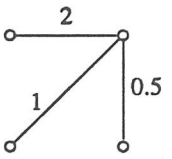

Figura 4.13: Pesos Locales Considerados en el Alineamiento Temporal Dinámico.

etiqueta	número de vectores por segmento	bits por código	*DE(dB)*	*DC*	*DISNG*
DTW-SegQ	2.23	10	3.57	3.29	0.75
DTW-Matrix	2.20	10	3.44	3.17	0.70
MQ	2.00	10	3.56	3.29	0.71
SegQ-RETARDO	2.39	10	3.55	3.27	0.73

Tabla 4.10: Resultados de la Segmentación con DTW, MQ y SegQ.

del sistema descrito a nuestra secuencia de test se obtienen los resultados que se muestran en la tabla (4.10). En la tabla (4.10) se ha etiquetado con *"DTW-SegQ"* al esquema para el que se realiza un DTW tomando como diccionario de patrones el diseñado para el esquema SegQ (apartado 4.7.2) y se ha etiquetado con *"DTW-Matrix"* al mismo procedimiento pero en este caso utilizando el diccionario diseñado para el MQ con $L = 3$, (ver apartado 4.7.1). Se extraen la siguientes conclusiones:

- Aun teniendo unas longitudes medias de segmento muy próximas, con el esquema *DTW-Matrix* se obtiene una distorsión menor que el *DTW-SegQ*. Lo cuál es atribuible al hecho de que, como ya se ha comentado, la secuencia de entrenamiento para el SegQ es insuficiente.

- Comparando con el MQ, se observa que se obtiene una mejor cuantización, en el sentido que para la longitud media de los segmentos obtenidos se consigue una distorsión menor que en el caso matricial.

- De la comparación del esquema *DTW-SegQ* con el *SegQ-RETARDO* (utilizan el mismo diccionario de patrones) se observa que, si bien consiguen unas distorsiones medias bastante próximas, el esquema *SegQ-RETARDO* obtiene una segmentación con duraciones medias superiores, lo que se traduce en un *bit-rate* medio menor. La razón de esta diferencia puede estribar en el hecho de que aunque el *DTW-SegQ* hace un alineamiento temporal dinámico, las decisiones son realizadas localmente, mientras que el *SegQ-RETARDO* realiza un casamiento estático pero la segmentación se realizada más globalmente dentro de los límites impuestos por el parámetro Δ (apartado 4.7.2). Y lo que es más, un problema evidente es que para el sistema DTW, además de codificar las longitudes de los segmentos que son variables, ha de codificarse el camino de alineación o alineamiento seguido para poder decodificar adecuadamente. Esta información lateral adicional supondría un incremento en el *bit-rate* que hace que no se justifique la utilización del esquema *DTW-SegQ*.

- Una de las fuentes de distorsión en el SegQ es el hecho de almacenar los segmentos código con una longitud fija. Es obvio que si se almacenaran múltiples diccionarios cada uno de ellos con distintas longitudes para los segmentos se mejoraría en calidad; ahora bien, el precio a pagar sería una mayor necesidad de memoria (tanto en el codificador como en el decodificador) y por supuesto se necesitaría una secuencia de entrenamiento aun mayor. En [Peter90] se evalúa tanto objetivamente como subjetivamente la incorporación de esta sencilla idea.

- Otra de las fuentes de distorsión en el esquema SegQ es la concatenación temporal abrupta de segmentos. El hecho de la coarticulación entre fonemas, sugiere que una suavización temporal de las concatenaciones mediante un solapamiento gradual entre segmentos adyacentes pueda mejorar la calidad resultante [Honda92].

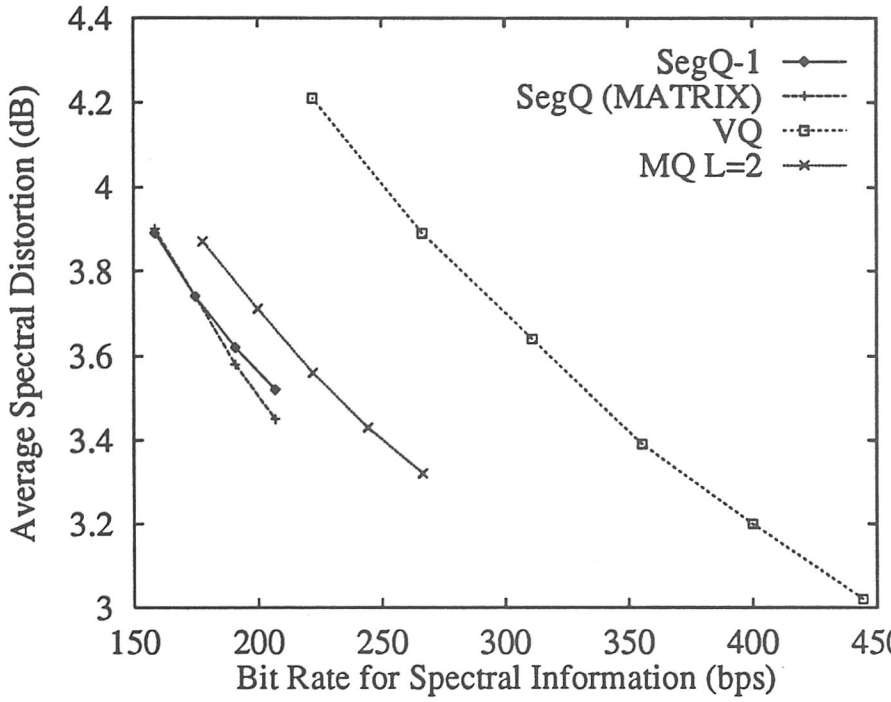

Figura 4.14: Cuantización con SegQ, SegQ MATRIX, MQ y VQ.

- Una limitación fundamental en el SegQ es la necesidad de una secuencia de entrenamiento muy grande. En particular, con los umbrales elegidos, el número de segmentos de entrenamiento nunca fué superior a 30000 vectores. Dicha secuencias son claramente insuficientes cuando se quiere caracterizar adecuadamente la estadística de la fuente. En la gráfica (4.14) se evidencia esta limitación, para lo cuál se representa el esquema etiquetado *SegQ-1*, junto con un esquema (etiquetado SegQ MATRIX) que utilizando el mismo procedimiento de codificación utiliza como diccionario el diseñado para el cuantizador matricial MQ L=3. A la vista de la gráfica (4.14), la limitación comentada se pone más de manifiesto cuanto mayor es el tamaño del diccionario, es decir cuanto mayor es el bit-rate.

  Motivados por esta limitación, para soslayar este problema, en [Honda92] se propone el diseño del diccionario SegQ para un solo locutor (secuencia de entrenamiento monolocutor menos exigente en cuanto a su tamaño) y posterior-

mente, en la fase de codificación, se propone realizar una cuantización adaptable actualizando los centros de los diccionarios atendiendo a la estadística del locutor actual.

4.8 Cuantización Multi-Trama

Otro procedimiento experimental para conseguir una aproximación a la curva R(D), puede encontrarse en el trabajo [Kemp91], donde se propone el sistema de codificación multi-trama (*"Multi-Frame Coding" (MF)*). La idea básica consiste en aprovechar las dependencias existentes entre vectores adyacentes (correlación inter-trama) y, para ello, codificar sólo algunos de los vectores de la secuencia de entrada, regenerando en el decodificador los no codificados mediante un sencillo procedimiento de interpolación lineal.

En dicho trabajo se propone considerar bloques de entrada de 8 vectores, que supone un retardo de 180 ms, y de entre los 8 vectores considerados, codificar solo cuatro de ellos. De entre todas las posibles, se eligen aquellas cuatro posiciones o vectores dentro del bloque que impliquen una distorsión promedio por trama menor.

Para el desarrollo del MF original se usan diccionarios producto (split-VQ, apartado 4.5.3) de 18 bits para caracterizar cada uno de los vectores elegidos en el bloque, 9 bits se emplean para codificar el subvector formado por los 4 primeros coeficientes LSP y los otros 9 bits se emplean en codificar los 6 coeficientes restantes. Este tamaño de diccionarios implica que para caracterizar la información espectral de un bloque (8 vectores) se emplean 9 bits por vector. Evidentemente, es necesario determinar cuáles de los 8 vectores se codifican, lo que se traduce en la necesidad de transmitir información lateral adicional. Para ello, se elige el siguiente procedimiento:

- Siempre se transmite información correspondiente al último vector del bloque.
- De entre las 35 posibilidades restantes, se descartan arbitrariamente 3, con lo que para este caso, sólo se necesitan transmitir 5 bits adicionales como

etiqueta	bits por vector	*DE(dB)*	*DC*	*DISNG*
MF-Split	9.625	2.63	2.37	0.49
VQ	9	3.20	2.93	0.64
VQ	10	3.02	2.73	0.58

Tabla 4.11: Resultados del MF-Split y VQ.

información lateral por bloque para codificar las posiciones que ocupan los vectores elegidos.

Por lo tanto, para el MF de referencia (etiquetado MF-Split) se tendrá un *bit-rate* promedio de 9.625 bits por vector, que cuando es aplicado a nuestra secuencia de test se obtienen los resultados mostrados en la tabla (4.11). Para la obtención de dicha tabla se ha utilizado la distancia pesada IHM, tanto para el VQ como para el MF.

Finalmente, en la gráfica (4.15) se muestra el comportamiento del MF-Split para distintos tamaños del diccionario utilizado para codificar los vectores elegidos (variando desde 4+4 hasta 11+11 bits). En dicha gráfica, se incluyen también los resultados obtenidos al considerar un MF usando un diccionario VQ convencional (no split), etiquetado MF, considerando desde 5 hasta 12 bits para el diccionario, y por comparación también se incluyen los resultados del esquema etiquetado *SegQ-1* (apartado 4.7.2), que como ya se comentó supone un retardo tendente a infinito.

Observando la gráfica (4.15) se concluye que

- Para bit-rates inferiores a 350 bps, la utilización de un diccionario convencional en el esquema MF, consigue unos resultados mejores que el correspondiente MF-Split, donde se utiliza un diccionario por división. La razón evidente es que para este último se están desaprovechando las dependencias intra-trama. No obstante, cabe resaltar que con el esquema MF no se pueden obtener distorsiones menores a las mostradas, pues para ello el tamaño del diccionario se alejaría de lo que es experimentalmente razonable (más de 12 bits). Para

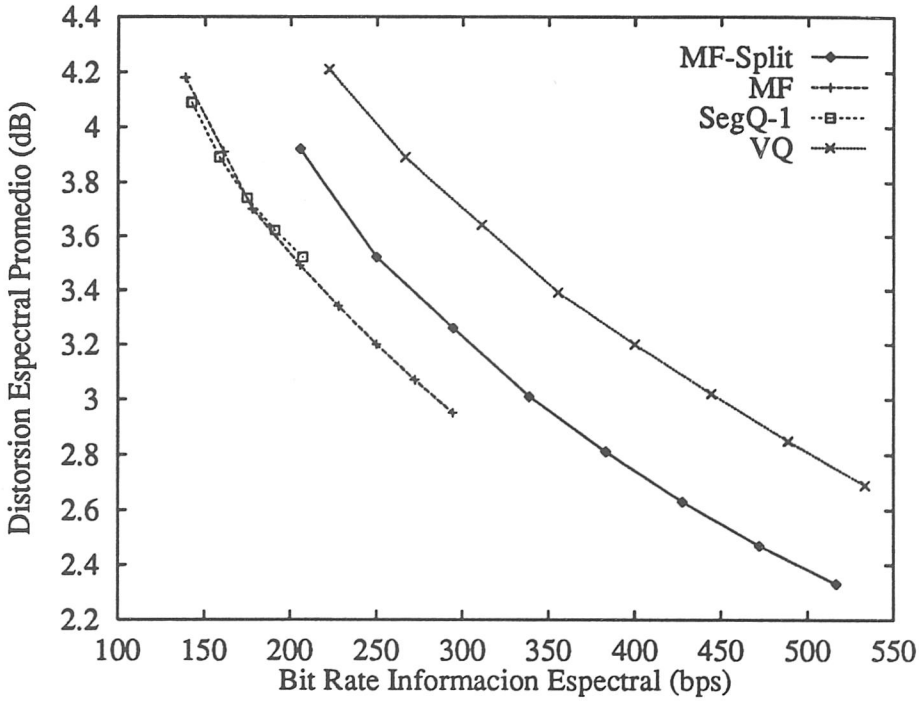

Figura 4.15: Cuantización con MF-Split, MF, SegQ y VQ.

obtener bit-rates mayores hay que recurrir a algún procedimiento subóptimo como es la utilización de un split-VQ, utilizada en el esquema MF-Split.

- De la comparación con el esquema SegQ, se concluye que haciendo uso exclusivamente de las dependencias inter-trama lineales, se obtienen unos resultados comparables, si no mejores, a los obtenidos con un esquema que intenta hacer uso de todas las dependencias estadísticas entre vectores. La razón de esta aparente contradición reside en que, debido a la falta de entrenamiento, los diccionarios SegQ no recogen adecuadamente la estadística de la fuente.

4.9 Cuantización e Interpolación Combinadas (CQI)

Para la utilización de las dependencias lineales entre vectores, una posible solución sería la generalización de la cuantización escalar óptima (apartado 4.3) al caso vectorial, que implicaría la eliminación de dichas dependencias, llevando a cabo una ge-

neralización multidimensional de la transformada de Karhunen-Loeve (KLT), transformando los vectores de entrada a un nuevo espacio vectorial decorrelacionado y aplicando una VQ en la que se hubiera realizado una asignación de bits óptima para el diccionario (por ejemplo con un procedimiento de *prunning-tree* [Chou89]). Una solución similar es la *Descomposición Temporal*, donde la trayectoria de entrada descrita en el espacio multidimensional de representación es aproximada como una combinación lineal de funciones base o eventos. Dicha formulación fué propuesta en [Atal83] y posteriormente ha sido desarrollada eficazmente en [Cheng91] y [Cheng93], entre otros. Ahora bien, aunque la KLT supone la solución óptima, la cantidad de cálculo necesario hace que dicha transformada sólo se use generalmente como límite teórico. Alternativamente se pueden usar transformadas que si bien son subóptimas, son menos exigentes computacionalmente, como por ejemplo la transformada discreta del coseno (DCT) bidimensional [Farvar89].

No obstante lo anterior, para hacer uso de las dependencias lineales, en este trabajo se toma como punto de partida el MF, y se propondrá un sistema que soslaye sus debilidades.

Como se ha mostrado en el apartado anterior, una forma de reducir el bit-rate es codificar algunos de los vectores de la secuencia de entrada, y el resto reconstruirlos mediante interpolación lineal. Uno de los problemas relacionados con la interpolación reside en localizar las posiciones donde hay un cambio espectral significativo en la trayectoria temporal descrita por los vectores LSP en el espacio multidimensional; a dichas posiciones y vectores correspondientes se les llamará *puntos de ruptura* y *vectores de ruptura* respectivamente.

En la región entre 200-300 bps, el esquema MF es de entre los aquí simulados el procedimiento que hasta ahora obtiene las mejores resultados (en el sentido de aproximar la curva $R(D)$). No obstante, dicho sistema impone unas limitaciones, que básicamente se pueden concretar en

- La búsqueda de los puntos de ruptura se realiza en bloques de longitud fija, es más, el último vector del bloque siempre es forzado a ser un vector de ruptura

- Además, en el sistema mencionado se supone que el número de puntos de ruptura en la trayectoria es fijo (cuatro) dentro de cada bloque

Es de esperar que eliminando estas limitaciones se obtendrá un sistema con mejores prestaciones que el MF de referencia. Para lo cual, se propone el siguiente procedimiento para la codificación que se denomina Cuantización e Interpolación Combinadas (*CQI: "Combined Quantization Interpolation"*) [Lopez93a] con el cual se concluye el presente capítulo de este trabajo de investigación.

Sean x_i y $\hat{x}_i$ el vector de entrada de coeficientes LSP y su versión cuantizada en el instante de tiempo $t = t_i$. El objetivo de un codificador CQI consistirá en identificar los instantes de tiempo $\mathcal{J} \equiv \{t_{j1}, t_{j2}, \cdots\}$ y sus correspondientes vectores ruptura $\{\hat{x}_{j1}, \hat{x}_{j2}, \cdots\}$.

Posteriormente, en el decodificador, la trayectoria multidimensional original, será reconstruida mediante interpolación lineal entre los vectores correspondientes a los puntos de ruptura determinados. Concretamente, el vector reconstruido en $t = t_i$ será $\hat{x}_i$ si $t_i \in \mathcal{J}$ ó será $\hat{x}_i^*$ si $t_i \notin \mathcal{J}$, donde se ha notado con $\hat{x}_i^*$ al vector resultante de la interpolación lineal en el instante t_i entre los vectores correspondientes al anterior y posterior puntos de ruptura de la secuencia.

La información a transmitir por el codificador será pues el conjunto de índices que identifican a los vectores código correspondientes a los puntos de ruptura, junto con sus posiciones.

Para encontrar la secuencia $\mathcal{J} \equiv \{t_{j1}, t_{j2}, \cdots\}$, se exigirá que el error cuadrático medio pesado (IHM) entre el vector de entrada y el vector resultante de la interpolación esté por debajo de un umbral predeterminado. Los vectores de coeficientes LSP en los instantes de tiempo $t_{j1}, t_{j2}, \cdots$ se denominan *vectores de ruptura aceptables*. Por supuesto, de entre todos los posibles conjuntos $\mathcal{J}$ el objetivo es determinar aquél que implique un *bit-rate* mínimo, lo que corresponderá a encontrar el conjunto de vectores ruptura aceptables con la mayor separación promedio posible.

La idea básica que subyace en este procedimiento se resume en determinar el conjunto $\mathcal{J}$ de la siguiente manera

Dado un vector de ruptura aceptable, para encontrar el siguiente punto de ruptura, desplazar la posición hasta que la condición de aceptabilidad sea violada.

Basándose en este procedimiento, la decisión realizada es evidente que tiene un carácter remarcadamente local, en el sentido de que las posiciones de ruptura se determinan teniendo en cuenta sólo dos vectores ruptura adyacentes. Para realizar una interpolación más global (por supuesto con la penalización de tener una retardo mayor), una vez que se haya determinado la posición del siguiente vector de ruptura aceptable, se repetirá el mismo procedimiento para determinar un segundo vector de ruptura. Después de determinar los dos vectores de ruptura siguientes, el primero de ellos es desplazado en torno a su posición inicial de tal manera que se identifique cuál es la posición definitiva que minimice el error cuadrático medio pesado, evaluado para todos los vectores pertenecientes al bloque considerado. En este sentido, se puede considerar al procedimiento propuesto como un algoritmo *en dos pasadas.*

En el diagrama de flujo de la figura (4.16) se detalla el funcionamiento del algoritmo CQI, en el que con U y N se nota respectivamente el umbral del error cuadrático medio pesado y la separación máxima (en número de tramas) entre dos vectores ruptura adyacentes. La meta de cada iteración es la obtención de los puntos de ruptura de la secuencia, notados por el conjunto $\mathcal{J} \equiv \{t_{j1}, t_{j2}, \cdots\}$.

4.9.1 Resultados experimentales del algoritmo CQI. Medidas Objetivas.

Al aplicar el algoritmo de la figura (4.16) sobre la misma secuencia de test utilizada en todo el trabajo se obtienen los resultados mostrados en la gráfica (4.17). Como se ha indicado, es necesario determinar las posiciones de los vectores ruptura o, mejor dicho, la separación en número de vectores entre los puntos de ruptura, mediante información lateral que ha de transmitirse además de la información espectral propiamente dicha. Dichas separaciones se codifican con un código de longitud variable

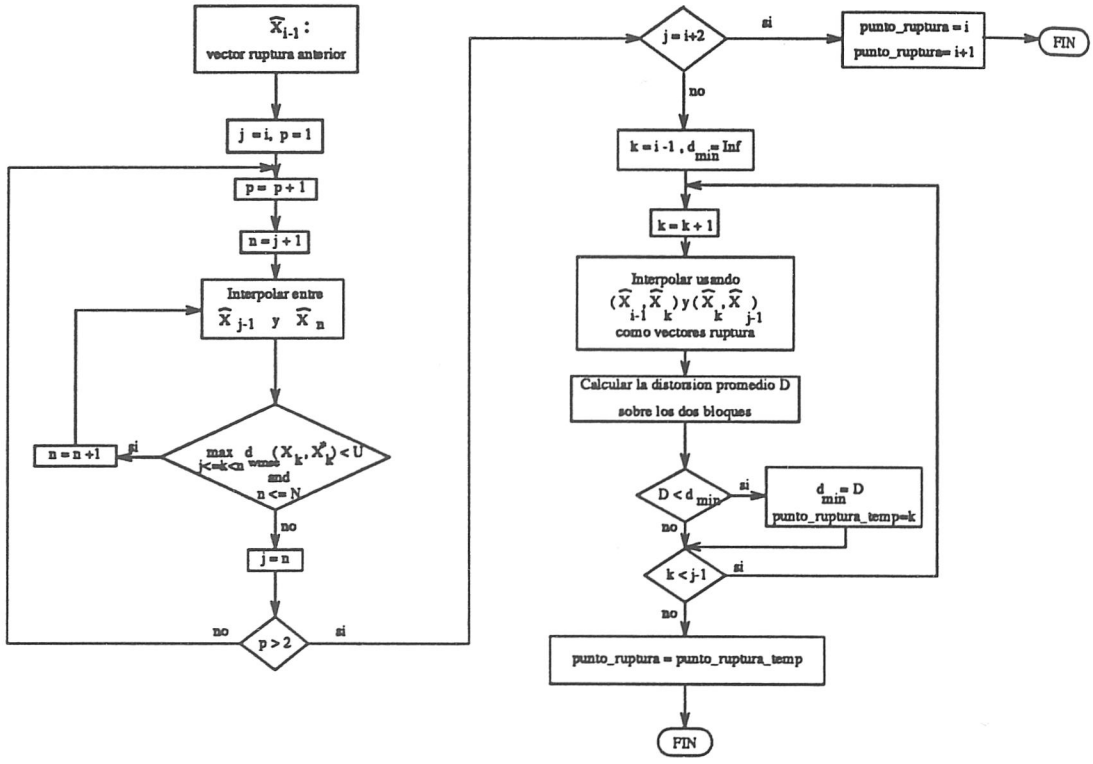

Figura 4.16: Diagrama de Flujo del Algoritmo CQI.

de Huffman. Por tanto, para la gráfica citada (4.17), el *bit-rate* espectral total se calcula teniendo en cuenta el tamaño en bits del diccionario usado para codificar los vectores ruptura más la información lateral resultante de codificar las posiciones.

En la gráfica (4.17) se incluye el resultado de dos esquemas CQI, etiquetados *CQI-10* y *CQI-12*, aludiendo al tamaño (en número de bits) del diccionario usado. Para cada una de las curvas, los distintos puntos corresponden a considerar distintos umbrales U, fijado un tamaño del diccionario. Por comparación, también se incluyen los resultados obtenidos anteriormente etiquetados con *MF-Split, MF y VQ*, donde para todos ellos se ha utilizado el error cuadrático medio pesado (IHM), como medida de distorsión.

Del experimento llevado a cabo se extraen las siguientes conclusiones,

- El CQI se comporta mejor que la simulación del esquema MF para bit-rates por debajo de 250 bps, caso de utilizar un diccionario de 10 bits y por debajo

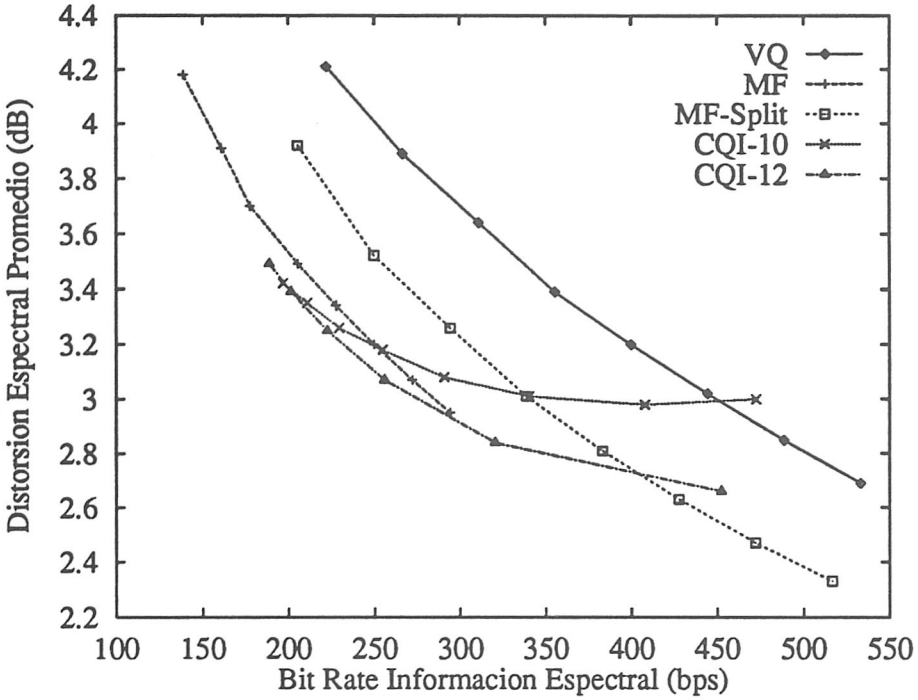

Figura 4.17: Algoritmo CQI junto a MF-Split, MF y VQ.

de aproximadamente 300 bps cuando el diccionario se codifica con 12 bits.

- La distorsión menor que se puede conseguir con el algoritmo CQI está determinada por el tamaño del diccionario considerado. Nótese que el algoritmo tiene una tendencia a la saturación al considerar umbrales menores, cuyo límite debe ser la distorsión obtenida con el VQ codificando todos los vectores.

- Para la obtención de menores distorsiones, es evidente que habría que utilizar diccionarios de mayor tamaño, para lo cuál se podría considerar el mismo algoritmo pero considerando split-VQ.

- Aunque no se ha incluido en la gráfica (4.17) por sencillez, observando la gráfica (4.15), el algoritmo CQI también se compara favorablemente con el etiquetado SegQ, siendo digno de mención el hecho de que en el CQI no es necesario un procedimiento de segmentación, lo que supone un menor coste computacional tanto en la fase de diseño como en la de operación.

- Aunque la distorsión obtenida en el esquema VQ con 10 bits es el límite al que debe tender el esquema CQI-10 al disminuir el umbral U, curiosamente, con dicho esquema se obtiene una distorsión ligeramente menor que la obtenida con el esquema VQ para los puntos de la gráfica con *bit-rates* superiores a 300 bps. La justificación de este hecho reside en que al incorporar la interpolación al proceso de cuantización, el número de posibles vectores código es 2^{2B} para cada una de las posibles posiciones del siguiente punto de ruptura. Es decir, introducir interpolación equivale colateralmente a incrementar el tamaño del dicionario, lo que se traduce en un comportamiento mejor.

Para realizar el experimento cuyos resultados se han recogido en la gráfica (4.17), se ha usado un valor de $N = 4$, lo que implica un retardo máximo para el codificador de 180 ms (correspondiente a la duración de 8 vetores de 22.5 ms). Cabría preguntarse cómo se comporta el sistema para distintos retardos o valores de N. En la gráfica (4.18) se presentan los resultados de considerar la aplicación del algoritmo CQI con distintos valores de N.

En cuanto al retardo considerado, observando la gráfica (4.18), se concluye que

- Cuando el valor de N es pequeño ($N = 2$), aun aumentando el valor del umbral, el sistema se queda atrapado en torno a los 250 bps. Ésto es debido a que no se permiten grandes separaciones entre vectores ruptura, con lo que el *bit-rate* no se puede reducir tanto como se desee.

- Al aumentar el valor de N, evidentemente para un umbral fijo U, se consigue un sistema con un *bit-rate* menor, ahora bien, a costa de un incremento en la distorsión introducida.

- Por debajo de 250 bps para la información espectral, el mejor esquema corresponde a considerar $N = 4$ (es decir 8 vectores), lo que quiere decir que no hay suficiente dependencia lineal entre vectores separados más de 180 ms, que justifique la utilización de retardos mayores.

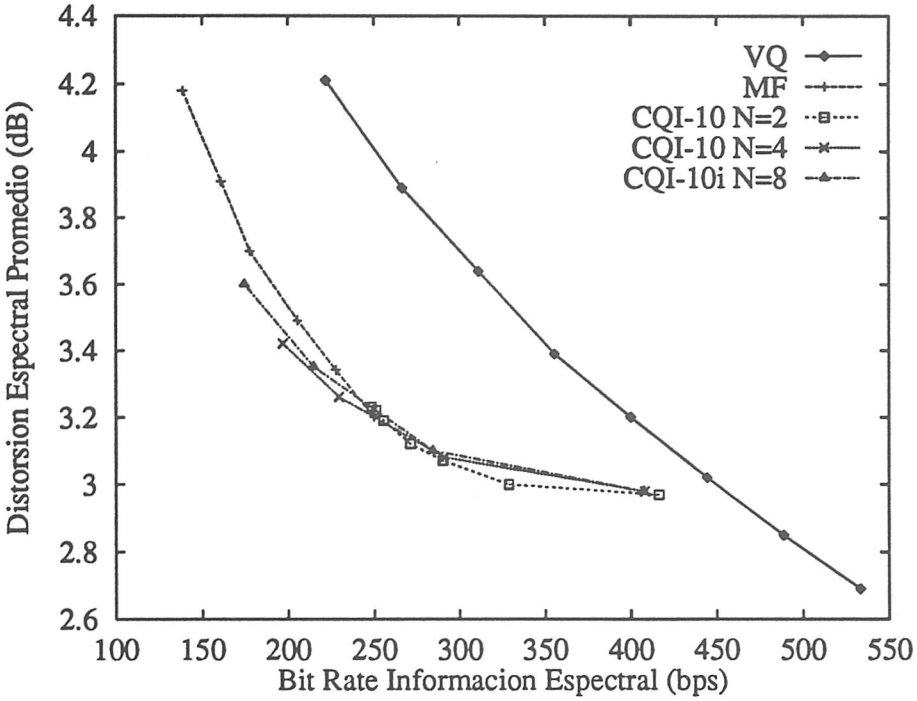

Figura 4.18: Algoritmo CQI para Distintos Retardos Máximos.

Para determinar la influencia de la utilización del código de Huffman (de longitud variable) al codificar la separación entre los vectores ruptura, en la gráfica (4.19) se representan los resultados obtenidos al codificar las posiciones con una codificación Huffman (etiquetado *CQI-12 N=4 Huff*) y los correspondientes de utilizar una codificación uniforme (etiquetada *CQI-12 N=4 Uni*) para caracterizar las posiciones de los vectores ruptura.

Como puede observarse en la gráfica (4.19), la mejora obtenida al usar codificación Huffman es más apreciable cuanto menor es el umbral, lo que indica que cuanto mayor es el umbral, las distintas posiciones de los vectores ruptura se van haciendo equiprobables, como cabría esperar.

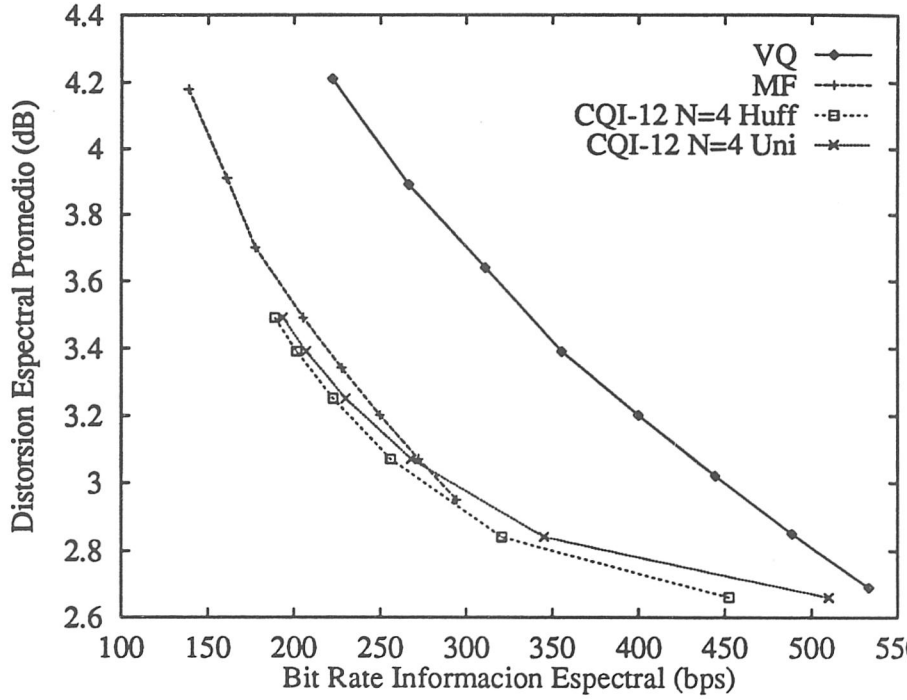

Figura 4.19: Influencia del Código Huffman en el Algoritmo CQI para la Codificación de las Posiciones de los Vectores Ruptura.

4.9.2 Resultados experimentales del algoritmo CQI. Medidas Subjetivas.

Para completar la evaluación del algoritmo CQI, se ha realizado un "rhyme test" de inteligibilidad (apartados 3.2.1 y 3.4). Las sílabas (consonante-vocal-consonante) se extrajeron de frases no pertenecientes al corpus de entrenamiento de TIMIT. Debido a la imposibilidad de crear un corpus específico de sílabas aisladas, en los resultados mostrados hay que tener en cuenta que las sílabas están afectadas por la coarticulación con los fonemas adyacentes.

En la tabla (4.12) se muestran los porcentajes de inteligibilidad obtenidos. Para el esquema CQI-10 se han evaluado dos esquemas con diferentes *bit-rates*, resultado de considerar distintos umbrales U. Se extraen las siguientes conclusiones:

- El bajo porcentaje obtenido para la señal original se debe, como ya se ha comentado, a la perturbación introducida por los efectos coarticulatorios.

	Original (4 kHz, 16 bits)	VQ 10 bits	CQI 10 bits	CQI 10 bits
bit-rate (bps)	128 k	444.4	357.2	254.4
Inteligibilidad	59.4%	37.3%	40.1%	33.3%

Tabla 4.12: Test Subjetivo para los esquemas CQI y VQ.

- El porcentaje obtenido con el esquema VQ supone una inteligibilidad relativa del 62.3% respecto a la original, y coherentemente con los resultados objetivos, al combinar la interpolación lineal junto con la cuantización, en el esquema CQI a 357.2 bps se obtiene una inteligibilidad relativa del 67.5%, incremento que puede ser justificado por idénticos argumentos a los dados para los resultados obtenidos en la evaluación objetiva (gráfica 4.17).

- Como cabría esperar, al aumentar el umbral U (es decir al reducir el bit-rate) la inteligibilidad se reduce significativamente, en concreto se obtiene una inteligibilidad relativa del 56.0% respecto a la original.

4.9.3 Comportamiento del algoritmo CQI en presencia de errores en el canal.

Para concluir con los resultados experimentales obtenidos con el algoritmo CQI, aun no siendo uno de los objetivos del presente trabajo de investigación, dado que este algoritmo es una de las contribuciones principales, se concluye este apartado caracterizando el comportamiento del algoritmo CQI cuando se considera un canal con errores.

Es evidente que si las posiciones de los vectores ruptura se codifican usando un código Huffman, cuando se incorpore información adicional redundante para corregir errores se ha de hacer especial hincapié en proteger dichos bits, ya que un simple error aquí resultaría en una desincronización entre el emisor y receptor para todo el resto de la secuencia, introduciendo una degradación significativa.

Sin embargo, cuando las posiciones de los vectores ruptura se codifican unifor-

	ϵ=0.0	ϵ=0.005	ϵ=0.01	ϵ=0.05	ϵ=0.1
U	DE	DE	DE	DE	DE
40	2.98	3.15	3.33	4.51	5.70
80	3.08	3.24	3.40	4.47	5.59
120	3.26	3.41	3.56	4.58	5.64
160	3.42	3.57	3.71	4.67	5.69

Tabla 4.13: Resultados del esquema CQI-10 para Diferentes Valores de ϵ y U.

memente (es decir, con un código de longitud fija), en principio no está claro cómo se verá afectada la calidad de la voz resultante en presencia de errores del canal. Suponiendo un canal binario simétrico, caracterizado por una tasa de errores denotada por ϵ, el esquema CQI-10 se ha simulado para el caso de N=4 con distintos umbrales U. Para la simulación realizada, se ha supuesto que el ruido afecta sólo al índice que caracteriza al vector código, y no afecta a las posiciones de los vectores ruptura. Como puede verse en la tabla (4.13), la degradación introducida para este caso es razonablemente aceptable, lo que justifica el hecho de que la mayor parte del esfuerzo en preservar la información de los posibles errores en el canal, debe ser realizado sobre los bits relacionados con las posiciones de los vectores de ruptura.

Por supuesto, dado que los vectores ruptura están codificados usando un VQ convencional, para preservar la información relativa a los índices, serán aplicables todas las técnicas y procedimientos desarrollados a este respecto para hacer una asignación de índices que minimice la distorsión para un canal dado, como por ejemplo los esquemas propuestos en [Farvar90], [Farvar93].

4.9.4 Posibles modificaciones al algoritmo CQI.

En un intento de mejorar las prestaciones del algoritmo CQI, en este subapartado se presentan posibles modificaciones de dicho algoritmo. La primera de ellas esta motivada por el hecho de que

> *dado un punto de ruptura, no siempre (o no necesariamente) el vector código vecino más próximo al vector de entrada será el mejor vector de*

ruptura

En otras palabras, observando el segundo bucle del diagrama de flujo de la figura (4.16), donde se determina $\hat{x}_k$, el vector ruptura en el instante $t = t_k$ no es necesariamente $q(x_k)$.

Por tanto la modificación planteada consiste en buscar tanto el instante de tiempo $t = t_k$ como el índice m que minimiza

$$D(i-1, k_m, j) \quad \text{with} \quad i \leq k \leq j, \; m \in \{0,1,2,3,r\} \tag{4.36}$$

donde $D(i-1, k_m, j)$ es el error cuadrático medio pesado obtenido al usar $\hat{x}_{i-1}, \hat{x}_{k_m}, \hat{x}_j$ como vectores de ruptura, $\hat{x}_{k_m}$ es el (m+1)-ésimo vector código cercano (en el sentido de la distorsión usada) a $\hat{x}_k$, con $m \in \{0,1,2,3\}$ y por último $\hat{x}_{k_r}$ representa al vector código más cercano al vector obtenido en el instante $t = t_k$ tras realizar una regresión lineal entre los vectores $\{x_i, \cdots, x_k\}$, forzando a que $\hat{x}_{i-1}$ pertenezca a la recta de regresión.

De esta forma, los procedimientos de cuantización e interpolación estarán más fuertemente acoplados, por lo que la modificación propuesta se etiqueta con el acrónimo SCQI ("*Strongly Combined Quantization Interpolation*").

En la gráfica (4.20) se muestran los resultados obtenidos al aplicar la modificación SCQI del algoritmo CQI inicialmente desarrollado. Para la obtención de dicha gráfica se optó por un diccionario de 12 bits y se tomó $N = 4$. En la tabla (4.14), se muestra el porcentaje resultante de seleccionar $\hat{x}_{k_m}$, con $m \in [0,1,2,3,r]$, como vector ruptura cuando la separación entre vectores ruptura es mayor que uno.

De los resultados mostrados en la gráfica (4.20) y en la tabla (4.14) se puede concluir que

- Con la modificación SCQI se obtienen mejores resultados que con el esquema CQI, ahora bien, las mejoras introducidas son tan poco significativas que por sí solas no justifican el incremento en coste computacional que el algoritmo

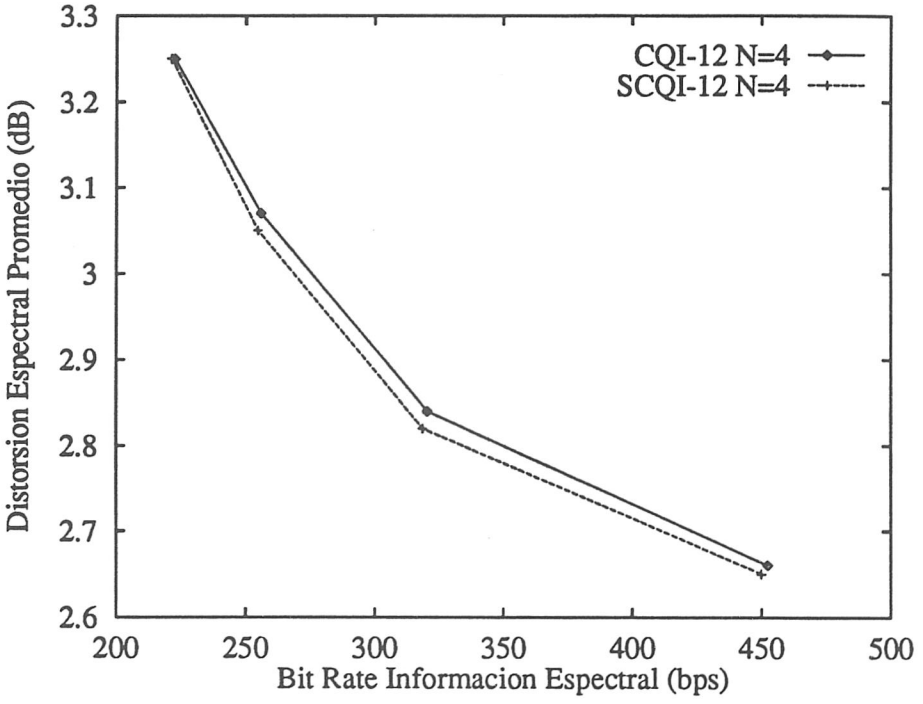

Figura 4.20: Resultados del SCQI-12 y CQI-12.

U	$\hat{\mathbf{x}}_{k_0}$	$\hat{\mathbf{x}}_{k_1}$	$\hat{\mathbf{x}}_{k_2}$	$\hat{\mathbf{x}}_{k_3}$	$\hat{\mathbf{x}}_{k_r}$
40	82.4 %	2.6 %	0.9 %	0.3 %	13.5 %
80	68.5 %	5.5 %	2.5 %	1.3 %	21.9 %
120	58.8 %	7.4 %	4.2 %	2.8 %	26.6 %
160	53.1 %	8.3 %	5.0 %	3.7 %	29.7 %

Tabla 4.14: Porcentaje de Elegir $\hat{x}_{k_m}$ como Vector Ruptura para el SCQI.

SCQI exige.

- Cuanto mayor es el umbral U (mayor es la separación media entre vectores ruptura), mayor es el porcentaje de elegir $\hat{\mathbf{x}}_{k_r}$ como vector ruptura, lo cual es lógico, pues cuanto mayor es el número de vectores para hacer la regresión es preferible utilizar como vector ruptura el vecino más próximo al vector resultante de la regresión, que el vector original.

Otra modificación consistió en considerar una generalización a un orden superior de la interpolación realizada. En particular, se aplicó el método de mínimos cua-

drados realizando una regresión de segundo orden, para ajustar la mejor parábola multidimensional que contenga los tres vectores formados por el vector ruptura anterior y los dos candidatos siguientes, evaluados exhaustivamente para todas las posibles combinaciones, dado un retardo máximo prefijado. Los resultados obtenidos fueron similares a los obtenidos para el caso lineal, indicando que la dependencia entre vectores es de primer orden, haciendo innecesaria la complejidad introducida al considerar dependencias de orden superior.

Adicionalmente, en un intento de eliminar toda la redundancia posible entre los vectores de ruptura, se realizó un experimento consistente en codificar los vectores ruptura utilizando una máquina de estados finitos determinista, en la que la función del estado siguiente (que proporciona qué diccionario usar) depende del estado (es decir, del diccionario anterior elegido) y del vector código anterior (en ese sentido se puede interpretar como una adaptación *backward*). Se consideró una VQ de estados finitos (*FSVQ: Finite-State Vector Quantization*) con velocidad de bits por segundo variable [Hussa93], donde no todos los diccionarios (o estados) tienen el mismo tamaño. Las mejoras introducidas no fueron significativas, indicando que poca dependencia o redundancia cabe esperar entre los vectores ruptura adyacentes, reforzando el hecho de que fijado un retardo máximo, con el algoritmo CQI se aproxima experimentalmente de la mejor forma posible la curva $R(D)$, objetivo de este trabajo de investigación.

Capítulo 5

Codificación de la Excitación

En este capítulo se va a resumir toda la investigación realizada relativa al estudio de la codificación de la excitación o estructura fina del espectro. En el capítulo 2 de esta memoria se introdujeron conceptos relativos al análisis de la excitación, así como un resumen de los estandares federales FS-1015 y FS-1016. Aquí se presentará un resumen de las aportaciones relativas a la codificación de la información contenida en la estructura fina del espectro.

5.1 Introducción

Tradicionalmente, cuando se trata de reducir el *bit-rate* al codificar la señal de excitación, se opta por un modelo simplificado que consiste en clasificar el residuo para cada trama de análisis de la información espectral tomando una decisión binaria: sonoro o no_sonoro. Esto da lugar a un codificador totalmente paramétrico para el que, además de codificar la información espectral, se requiere transmitir algún parámetro directamente relacionado con la energía de la señal junto con el periodo fundamental, caso de tratarse de una trama sonora.

Aceptando las limitaciones (ya expresadas en un capítulo anterior) del modelo paramétrico (vocoder), aquí se asumirá dicha drástica simplificación y se propondrán y compararán varios procedimientos para codificar dicha información.

La parametrización elegida es coherente con el objetivo global del presente trabajo, encontrar el mejor compromiso rate-distorsión en la región de muy bajo *bit-rate*, donde es conocido que los codificadores híbridos (y tras varias pruebas informales subjetivas así se ha corroborado) sufren una degradación significativa en la calidad.

Las técnicas aquí presentadas serán aplicables siempre que se trate de un codificador paramétrico, independientemente de si se trata de un análisis en bucle cerrado o en bucle abierto.

5.2 Cuantización de la Energía

Para caracterizar la energía de la trama, se adoptará como parámetro el logaritmo de la potencia de la secuencia error de predicción, definida por

$$\alpha_N = \log \frac{\alpha}{N} \tag{5.1}$$

siendo α la energía de dicha secuencia definida en (2.6) y N el número de muestras de la ventana de análisis.

1	2	3	4	5	6	7	8	9	10
0.75	0.45	0.23	0.12	0.06	0.03	0.02	0.03	0.05	0.07

Tabla 5.1: Coeficientes de Autocovarianza Normalizada para α_N.

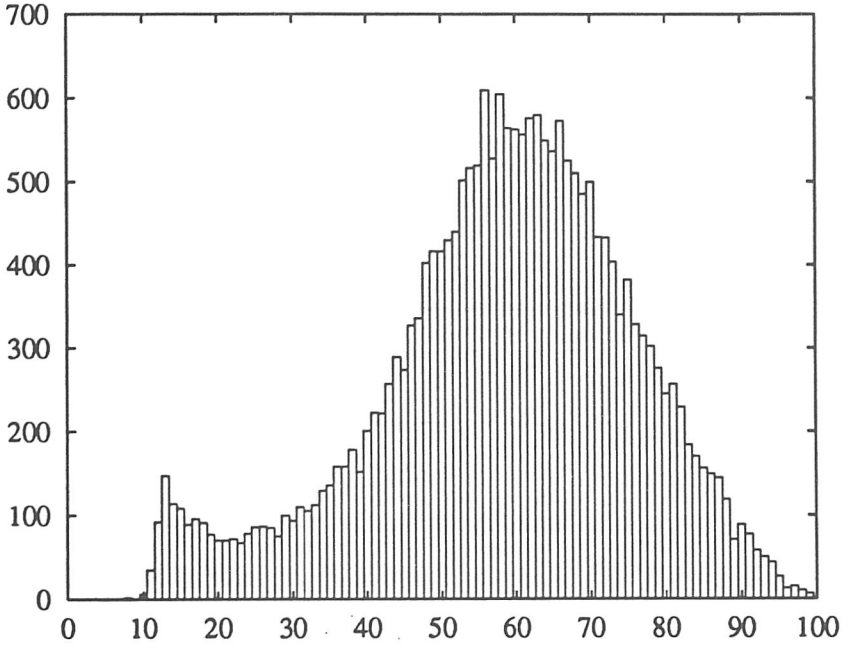

Figura 5.1: Histograma de α_N.

En la tabla (5.1) se presentan los valores de la autocovarianza normalizada de la secuencia de los α_N, para la secuencia de entrenamiento TIMIT-2, descrita en el apartado (3.5). Igualmente en la gráfica (5.1) se representa el histograma o distribución para la fuente generadora de la secuencia de α_N, extraida de la misma secuencia de entrenamiento TIMIT-2, usando las mismas condiciones de análisis de la tabla (2.6). El eje de abcisas de dicha gráfica corresponde al número del intervalo considerado. Siendo $\alpha_{Nmax} = 13.77$ y $\alpha_{Nmin} = -2.71$, para la secuencia considerada.

En los siguientes subapartados se presentan y simulan varios esquemas de codificación-decodificación para la secuencia de α_N. Encontrándose que el procedimiento

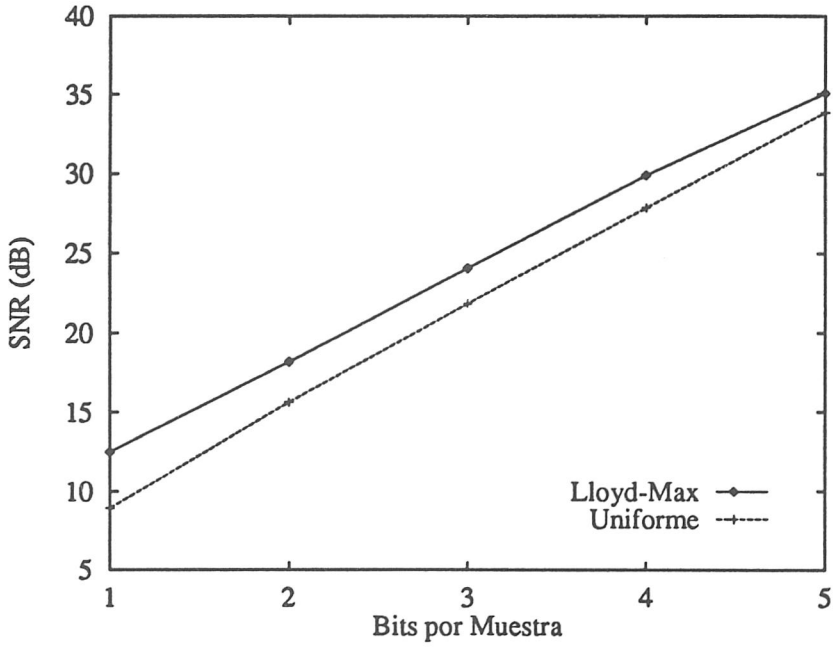

Figura 5.2: Cuantización uniforme y Lloyd-Max para α_N.

más adecuado de entre los simulados es de nuevo aquél que, como en el caso de la información espectral, combina la cuantización junto con la interpolación lineal.

5.2.1 Cuantización Lloyd-Max

Como punto de partida, en la gráfica (5.2), se presentan los resultados de codificar los α_N con un cuantizador uniforme, a la vez que se representan los resultados de cuantizar la misma secuencia de test con un cuantizador diseñado con un procedimiento Lloyd-Max. La medida de distorsión utilizada consiste en la SNR (expresión 3.4) y el *bit-rate* está expresado en bits por muestra.

5.2.2 Cuantización Diferencial

Particularizando el esquema DPCM vectorial (ver figuras 4.4 y 4.5) para el caso escalar al codificar los α_N (expresión 5.1) se obtuvieron los resultados mostrados en la gráfica (5.3). En dicha gráfica se han representado los resultados etiquetados por *DPCM-O* con $O = 1, 3, 5$, correspondientes a considerar distintos ordenes O para

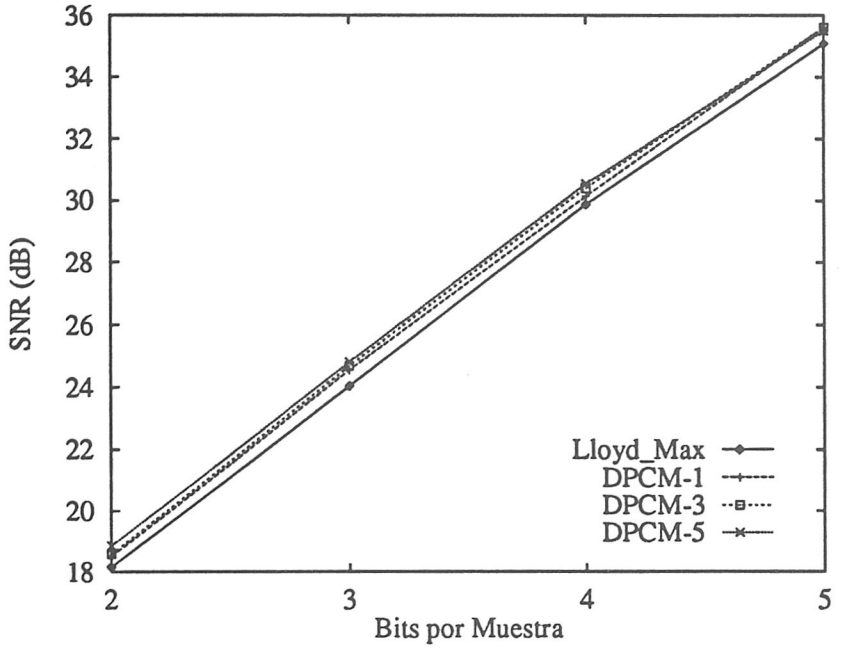

Figura 5.3: Cuantización DPCM y Lloyd-Max para α_N.

ORDEN O	a_1	a_2	a_3	a_4	a_5
1	0.96				
3	1.11	-0.33	0.18		
5	1.07	-0.28	0.05	0.01	0.11

Tabla 5.2: Coeficientes de Predicción para α_N, usando DPCM.

el predictor. Por comparación, también se incluye la curva obtenida en el apartado anterior, etiquetada *Lloyd-Max*. Para los esquemas *DPCM-O*, el cuantizador de diferencias se ha diseñado igualmente con un procedimiento Lloyd-Max monodimensional sobre las diferencias obtenidas aplicando el predictor a la secuencia de entrenamiento TIMIT-2 en bucle abierto. En la tabla (5.2) se representan los coeficientes de predicción utilizados para los distintos ordenes O obtenidos mediante el método de autocorrelación con un procedimiento *Levinson-Durbin* [Rab78] de minimización de la energía residual. A la vista de los resultados obtenidos se observa que

- cuando el número de bits por muestra es menor que 5, hay un mejor comportamiento del esquema DPCM que el esquema etiquetado Lloyd-Max.

- El sistema tiende a la saturación al aumentar el orden de predicción (no se observaron mejoras significativas para ordenes de predicción superiores a 5), lo cual es lógico teniendo en cuenta los valores de autocorrelación normalizada estimados para la fuente en la tabla (5.1).

- Igualmente el sistema tiende a la saturación al aumentar el número de bits por muestra, como se observa en la figura en el punto correspondiente a 5 bits por muestra; ésto indica que a partir de este punto las diferentes ganancias de predicción (ver apartado (4.6)) para los distintos ordenes considerados no introducen variaciones significativas.

- Siempre es de esperar un mejor comportamiento realizando un diseño en bucle cerrado y, por supuesto, iterando el diseño del predictor con el diseño del cuantizador como los procedimientos descritos para el caso vectorial.

- Claramente, uno de los puntos débiles del esquema DPCM, es que el predictor se ha diseñado teniendo en cuenta la estadística de retardo largo de la fuente. En otras palabras, en el diseño realizado se está suponiendo que la fuente exhibe un carácter estacionario. Es de esperar que un esquema que se adapte a la estadística de retardo corto, (como el esquema ADPCM (*"Adaptive DPCM"*), utilizado y aprobado en [CCITT90] para transmisión digital a 40, 32, 24 y 16 kpbs) proporcione mejores resultados.

5.2.3 Modulación usando Codificación por "Trellis" (TCM)

Teniendo presente la dualidad entre la modulación en comunicaciones digitales con el problema de la codificación de la fuente, recientemente [Marcell90] se han propuesto sistemas de codificación basados en los desarrollos que con gran éxito Ungerboeck propuso [Unger82] para mejorar el comportamiento de los sistemas de transmisión

digital en presencia de ruido sin sacrificar velocidad de transmisión ni exigir más ancho de banda. En este subapartado se resumen las ideas básicas de la modulación usando *"trellis"*, que servirán como punto de partida para introducir de una manera dual los *"trellis"* aplicados a la codificación de la fuente.

Tradicionalmente, en los sistemas de comunicación digital se realizan dos procesos independientes que consisten por un lado en codificar introduciendo redundancia para corregir posibles errores del canal y posteriormente realizar la modulación (selección y transmisión del correspondiente punto de la constelación).

La aportación básica que subyace en TCM consiste en realizar conjuntamente la modulación junto con el proceso de codificación del canal, usando el procedimiento literalmente llamado *"mapping by set partitionning"* [Unger87a]-[Unger87b]. Consiste en dividir el espacio de representación o constelación en subconjuntos, tal que dichos subconjuntos (o estados) maximicen la distancia promedio evaluada para los puntos pertenecientes a dicho subconjunto. A la vez, se imponen ciertas restricciones a las posibles secuencias generadas, limitando la posible secuencia de estados en el codificador. Las implicaciones de esta nueva aproximación, llamada *"Trellis" Coded Modulation (TCM)*, se resumen en

1. El número de puntos usados en la constelación es mayor del que sería necesario para conseguir la velocidad de transmisión requerida, con ello se introduce redundancia para llevar a cabo un corrección de errores *hacia adelante.*

2. Las posibles secuencias generadas por el codificar se limitan usando un código convolucional que consiste en una máquina de estados finitos, a la que debido a sus propiedades estructurales se le asocia una estructura en rejilla o *"trellis"*, donde cada punto tiene asociado un estado y cada transición entre estados supone la elección de un determinado subconjunto para la siguiente decodificación.

En cuanto a la decodificación, es conocido que para los códigos convolucionales, suponiendo un canal con Ruido Gausiano Blanco y Aditivo (RGBA), el decodificador

de *máxima-probabilidad* se reduce al decodificador de *distancia-mínima* (Euclídea cuadrática), es decir se elige la secuencia de estados (o camino en el *"trellis"*) que minimiza la distancia a la secuencia de muestras de entrada.

Basados en este principio, el algoritmo de Viterbi [Forn73], (utilizado ya en este trabajo (apartado 4.7.2)), lleva a cabo esta tarea en la que para cada etapa del "trellis" y para cada uno de los estados va eligiendo como caminos supervivientes a los que minimizan la métrica asociada.

Una desventaja del Algoritmo de Viterbi es que el número de operaciones para decodificar un bit está relacionado exponencialmente con la *memoria* del codificador, lo que impone una limitación práctica en cuanto a la memoria máxima del código. No obstante, hay soluciones subóptimas, que permiten incrementar la memoria del código (equivalentemente incrementar la capacidad de corrección de errores) sin necesidad de evaluar todos los posibles caminos. Dichas técnicas son llamadas globalmente *decodificación secuencial* [Proak89], que aquí no se consideran por estar fuera del propósito de este trabajo de investigación.

5.2.4 Cuantización usando Codificación por "Trellis" (TCQ)

La codificación de la fuente usando "Trellis" ha sido un esquema ampliamente estudiado desde el punto de vista de la *teoría de la rate-distorsión.* Ahora bien, desafortunadamente la mayoría de los desarrollos probados son *no-constructivos*, lo que ha forzado tradicionalmente a diseñar los "trellis" de una manera subóptima.

Una de las aportaciones definitivas de la cuantización usando codificación por "trellis", es proporcionar un procedimiento constructivo para diseñar los "trellis", el cual se basa en las nociones de *"mapping by set partitionning"* de Ungerboeck para TCM, resultando en sistemas [Marcell90] de prestaciones comparables si no superiores a los "trellis" tradicionales.

Para codificar una fuente (en este caso la secuencia de α_N, expresión (5.1)), con R bits por muestra, usando un "trellis" AM de Ungerboeck, se tomará como alfabeto de salida al conjunto de puntos obtenido con un diseño Lloyd-Max de $(R+1)$

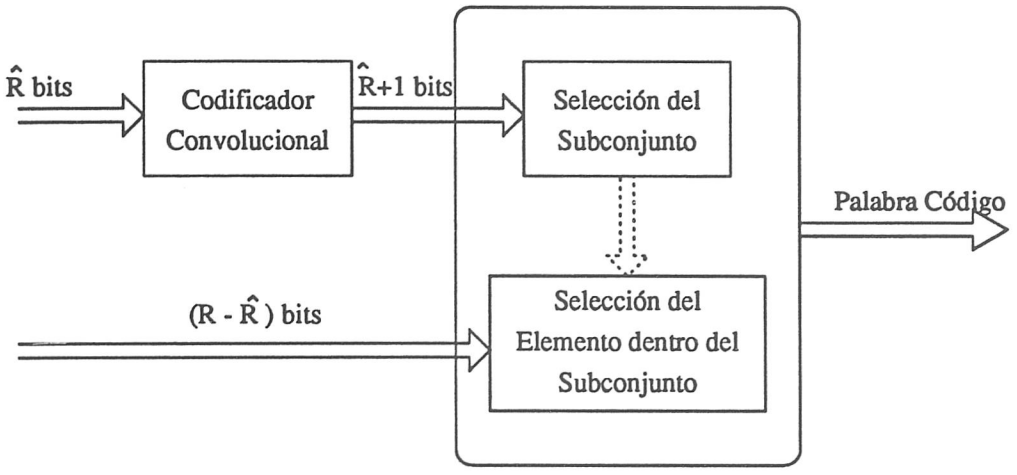

Figura 5.4: Diagrama del Cuantizador TCQ.

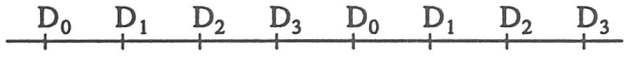

Figura 5.5: Etiquetado y partición de un TCQ con $R = 2$ y $\hat{R} = 1$.

bits por muestra. Estos puntos, son divididos en $S = 2^{(\hat{R}+1)}$ subconjuntos, con $\hat{R} \in I, \hat{R} \leq R$. $\hat{R}$ bits se expanden usando un código convolucional con un *rate* de $\hat{R}/(\hat{R}+1)$ (figura 5.4) y se usan para determinar el subconjunto al que pertenece el símbolo actual, y el resto $(R - \hat{R})$ se usan para identificar el punto de la constelación más cercano a la muestra de entrada dentro del subconjunto actual elegido. El etiquetado de los puntos de la constelación se realiza como siempre maximizando la distancia entre las posibles secuencias generadas por el codificador, de tal forma que para el caso AM, la asignación se realiza empezando por el más a la izquierda y etiquetando con la secuencia $D_0, D_1, \cdots, D_{S-1}, D_0, D_1, \cdots, D_{S-1}, \cdots$.

En la figura (5.5) se muestra el etiquetado realizado para el caso particular de $R = 2$, $\hat{R} = 1$, y en la figura (5.6) se muestra el "trellis" AM de Ungerboeck con $S = 4$, para el que se ha adoptado el convenio de etiquetar primero la transición superior y después la inferior con los correspondientes subconjuntos asociados. En la tabla (5.3), se muestran los resultados obtenidos sobre la misma secuencia de test (α_N), para $S = 4$ estados. La decodificación se hace mediante el algoritmo de

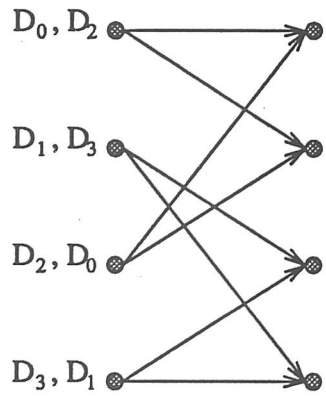

Figura 5.6: "Trellis" de Ungerboeck para $S = 4$.

P (tramas)	R (bits por muestra)	SNR (dB)
4	2	17.94
4	3	24.06
4	4	29.33
8	2	18.17
8	3	24.34
8	4	29.56

Tabla 5.3: TCQ con $S = 4$ para los α_n.

Viterbi, para bloques de profundidad P, lo que implica un retardo de P tramas. Como se observa en la tabla (5.3), al aumentar la profundidad de la búsqueda las prestaciones del sistema mejoran. Ahora bien, es claro que el sistema se puede mejorar realizando una búsqueda por cada muestra de entrada, es decir, en vez de utilizar un Viterbi que codifica todo un bloque, solapar dichas búsquedas y repetir el proceso muestra a muestra. Evidentemente el coste pagado por ello es un incremento en las necesidades de cálculo en un factor P. Los resultados de esta modificación se muestran en la tabla (5.4).

Hasta aquí se ha considerado tan solo el "trellis" de 4 estados. Evidentemente es posible considerar más estados, en concreto en la tabla (5.5) se muestran resultados para el caso de considerar 8 estados para un TCQ con solapamiento. En este caso

P (tramas)	R (bits por muestra)	SNR (dB)
8	2	18.54
8	3	24.60
8	4	29.73

Tabla 5.4: TCQ con Solapamiento $S = 4$ para los α_n.

P	S	R	SNR (dB)	S	R	SNR (dB)	S	R	SNR (dB)
8	4	2	18.77	8	2	18.68	16	2	17.03
8	4	3	24.76	8	3	23.76	16	3	21.99
8	4	4	29.75	8	4	28.63	16	4	-

Tabla 5.5: TCQ con Solapamiento para los α_n, Número de Estados 8.

se consideran las tres posibilidades mostradas en la figura (5.7), donde se muestran las estructuras y el etiquetado correspondiente a usar $S = 4, 8, 16$ subconjuntos.

Como se observa por comparación de las tablas (5.5) y (5.4), el aumentar el número de estados de 4 a 8 supone un incremento en las prestaciones del sistema (en términos de SNR), para todos los *rates* considerados, cuando el número de subconjuntos S es 4. Esta mejora es tanto mayor cuanto menor es el *bit-rate*. Para el caso de $S = 8$ subconjuntos, sólo se produce mejora cuando el número de bits por muestra es 2, y para el caso $S = 16$ no se produce mejora. Como conclusión, es evidente que las mejoras son tanto más significativas cuantos menos niveles de cuantización se consideran. Ésto se puede atribuir a que cuando el número de niveles aumenta, dado que el rango dinámico de la variable considerada es pequeño, el sistema tiende a la saturación, y poca mejora cabe esperar si se introducen más estados o más subconjuntos.

Finalmente, motivados por las mejoras introducidas al considerar un esquema diferencial (DPCM) ver gráfica (5.3), se va a desarrollar un procedimiento que apro-

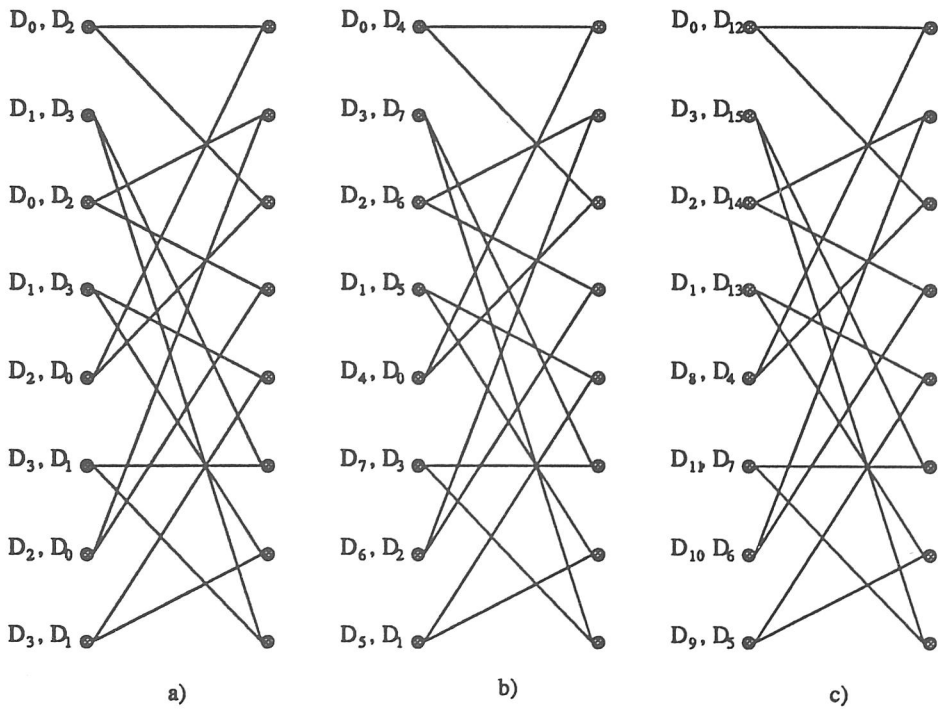

Figura 5.7: "Trellis" de Ungerboeck para 8 estados. a) $S = 4$. b) $S = 8$. c) $S = 16$.

veche la memoria que exhibe la fuente, incorporando un esquema predictivo al TCQ. En este caso, cada camino superviviente del "trellis", se puede usar para estimar una predicción de la muestra actual, y en este caso, codificar la diferencia entre la muestra predicha y la actual. Es de esperar que la varianza de la señal diferencia sea menor que la varianza de la fuente original, y que aunque la ganancia de cuantización sea menor, sea compensada con el incremento de la ganancia introducido en el bloque de predicción.

Toda vez que para un esquema diferencial se cumple que (ver expresión (4.13) y figura (4.4))

$$x_n - \hat{x}_n = e_n - \hat{e}_n \tag{5.2}$$

en el caso de un TCQ predictivo, para cada rama que emana de cada uno de los estados, se hará una cuantización escalar que determine el elemento del correspondiente subconjunto que minimice el error de cuantización $e_n - \hat{e}_n$. Esta distorsión será

acumulada al coste asociado a dicho camino, con lo cual cada camino superviviente tendrá asociada la secuencia de subconjuntos y el coste mínimo correspondiente de cuantizar la secuencia de entrada x_n. Al alcanzar una determinada profundidad P, de entre todos los estados se elegirá igualmente el de coste mínimo asociado, y se procederá recursivamente hacia atrás para recuperar la secuencia de subconjuntos asociada, realizando así una codificación de *distancia-mínima* que, como ya se ha dicho, corresponde al decodificador de *máxima-probabilidad.*

Más formalmente, utilizando la notación de [Marcell90], sea $\mathbf{x} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_P}\}$ la secuencia de muestras de entrada a codificar. Supóngase el *i-ésimo* instante de tiempo correspondiente a la muestra x_i de la secuencia, y que para cada nodo llegan 2 ramas y emergen otras 2. Sean

- N el número de estados
- *superviviente_en_k* el camino superviviente que termina en el nodo k en el instante de tiempo $t = i - 1$
- $\hat{x}^k_{i-j}$ con $k = 1, 2, \cdots, N$ los valores cuantizados correspondientes a la secuencia x_{i-j} con $j = 1, 2, 3, \cdots$ para el estado k
- $\hat{x}^k_{i|i-1}$ el valor predicho para la muestra actual x_i dado $\hat{x}^k_{i-j}$ con $j = 1, 2, 3, \cdots$ para cada estado k
- $d^k_i = (x_i - \hat{x}^k_{i-j})$ el residuo o error de predicción asociado con el *superviviente_en_k*
- $\rho_{i-1}(\mathbf{x}, \hat{\mathbf{x}}^k)$ la distorsión asociada con el *superviviente_en_k*
- D^k_l el subconjunto asociado con la rama que emana del nodo k y llega a l
- $\hat{D}^k_l$ el elemento perteneciente al subconjunto D^k_l más cercano a d^k_i
- Finalmente, para cada nodo l, sean $\hat{D}^{k_1}_l$ y $\hat{D}^{k_2}_l$ los dos elementos correspondientes a las ramas que unen los 2 nodos anteriores k_1 y k_2 al nodo l

N (estados)	S (subconjuntos)	R (bits por muestra)	SNR (dB)
4	4	2	20.25
4	4	3	25.84
4	4	4	30.23
8	4	2	20.44
8	4	3	25.96
8	4	4	30.36

Tabla 5.6: TCQ Predictivo con Solapamiento, $O = 8$ para los α_n.

Con la notación introducida, se tiene que

$$\rho_i(\mathbf{x}, \hat{\mathbf{x}}^l) = \min_{k \in \{k_1, k_2\}} (\rho_{i-1}(\mathbf{x}, \hat{\mathbf{x}}^k) + (d_i^k - \hat{D}_l^k)^2) \tag{5.3}$$

siendo

$$\hat{x}_i^l = x_{i|i-1}^{k'} + \hat{D}_l^{k'} \tag{5.4}$$

el correspondiente valor cuantizado para el instante de tiempo $t = i$, donde k' es el valor de k que minimiza la expresión (5.3). Este procedimiento es iterado desde $i = 1, 2, \cdots, P$, siendo P la profundidad considerada. En el caso de considerar solapamiento, tras decodificar la secuencia hacia atrás, se transmite el símbolo correspondiente a la muestra i, se incrementa i en 1, y se itera el procedimiento.

En la tabla (5.6) se muestran los resultados obtenidos tras la aplicación del esquema llamado *TCQ predictivo con solapamiento*, para 4 y 8 estados, con orden de predicción $O = 8$. Se ha considerado una profundidad de $P = 8$ y para el caso de 8 estados, basados en experiencias anteriores se elige $S = 4$.

Como conclusión final, en la tabla (5.6) se observa que el procedimiento mejor de entre los hasta ahora simulados para codificar la secuencia de α_N, resulta ser el TCQ predictivo con solapamiento para 8 estados con 4 subconjuntos. Evidentemente, es de esperar que con la incorporación de un esquema adaptable en la predicción, se obtenga una mejora adicional al esquema aquí simulado.

5.2.5 Cuantización e Interpolación Combinadas

Motivados por los excelentes resultados obtenidos para la codificación de la información espectral con el esquema CQI, ver apartado (4.9), aquí se aplicará dicho procedimiento particularizado para el caso escalar. Al considerar un retardo máximo igual a la duración de 8 tramas de análisis (es decir, 8 muestras) se obtienen los resultados mostrados en la gráfica (5.8). En dicha gráfica se etiqueta con *CQI-B* los resultados correspondientes al considerar un diccionario de B bits para codificar las *muestras ruptura*[1]. Los puntos para cada curva son el resultado de considerar distintos umbrales U (ver diagrama de flujo de la figura (4.16)). En la misma gráfica, para comparar, se han incluido tanto los resultados del esquema Lloyd-Max, junto con los mejores resultados obtenidos hasta ahora con el procedimiento TCQ predictivo con solapamiento $O = 8$, $S = 4$ etiquetado *TCQ_pred_solap*.

Tras los resultados mostrados en la gráfica (5.8) se concluye que

- Como puede observarse en torno a 2 bits por muestra el CQI ofrece unas prestaciones similares si no superiores que el mejor de los sistemas hasta ahora simulados, lo cual refuerza la idea de que combinando fuertemente los procesos de cuantización e interpolación se puede aproximar experimentalmente el límite teórico (en el sentido de la *función rate-distorsión*) que toda fuente tiene.

- Los mejores resultados se obtienen al utilizar el diccionario de 3 bits (CQI-3). Al igual que para la información espectral (gráfica 4.17), las dependencias lineales son tan fuertes que es incluso preferible (en el sentido de una mayor SNR) aproximar la trayectoria o evolución linealmente en el espacio de representación, que intentar cuantizar cada una de las muestras de dicha evolución. La razón de este comportamiento reside en que con el esquema CQI se están utilizando más niveles de cuantización, ya que no sólo son posibles los 2^B bits del diccionario utilizado sino también los posibles valores interpolados, que

[1] Se usa una terminología análoga a la introducida en el caso vectorial (apartado 4.9)

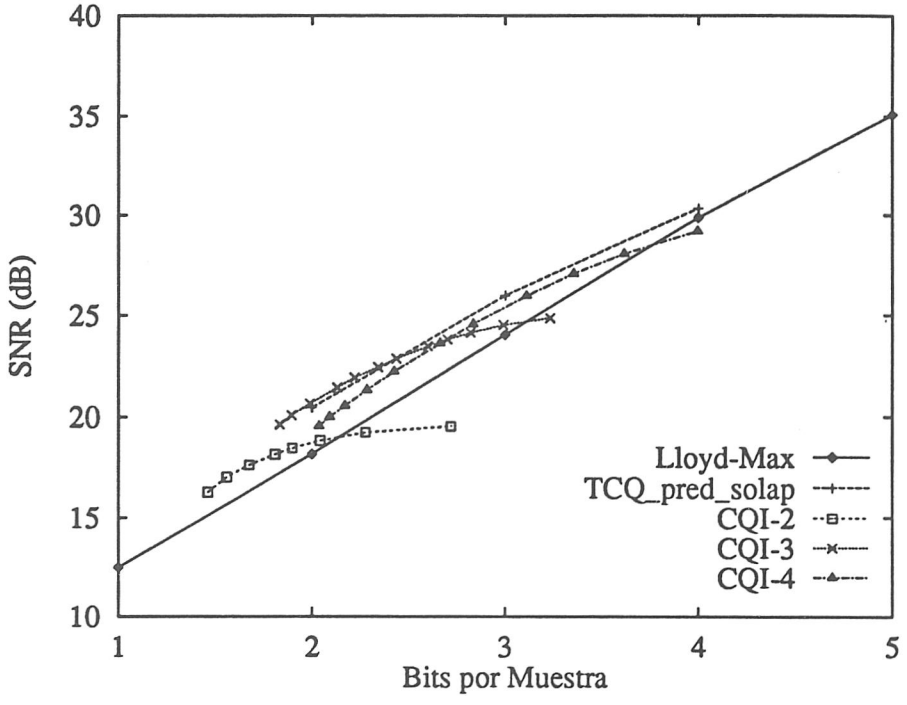

Figura 5.8: CQI Escalar, Lloyd-Max y TCQ Predictivo con Solapamiento para los α_N.

hacen un total de 2^{2B} para cada posición de la *muestra ruptura* siguiente.

- Para codificar las posiciones de los *puntos de ruptura* se ha utilizado una cuantización uniforme, luego es de esperar todavía mejores resultados si se codifican teniendo en cuenta la distribución particular de dicha fuente.

5.3 Cuantización de la Frecuencia Fundamental

Para acabar con la codificación de la excitación, y si se continúa suponiendo el modelo paramétrico, sólo resta aparte de la clasificación en sonoro/no-sonoro, codificar la frecuencia fundamental o pitch.

Para lo cual, en los siguiente subapartados se proponen distintos procedimientos para cuantizar dicha fuente. Se considera la misma secuencia de entrenamiento (TIMIT-2) y a su vez, para evaluar los distintos procedimientos se considera la

misma secuencia de test ("*core-test*" sugerida en la TIMIT).

La estimación de la frecuencia fundamental, junto con la clasificación sonoro/no-sonoro, no son un problema trivial. Son varias las causas que pueden inducir a error en la determinación del pitch: ya se ha citado que periodicidad no es sinónimo de sonoridad, además son posibles interacciones entre la cavidad resonante bucal y la excitación glotal, expresables en términos de correlaciones temporales, a la vez no siempre todas las tramas son fácilmente clasificables en sólo dos categorías, etc. Todas estas razones han justificado un gran esfuerzo por parte de la comunidad científica en este campo de investigación [Mark72],[Ross74],[Galvez91]. En el capítulo 2 se propuso un posible procedimiento en bucle cerrado para caracterizar la señal de excitación, si bien aquí se utilizará el procedimiento de análisis bien estudiado y eficazmente desarrollado en el estándar FS-1015 a 2400 bps [Kemp87].

Básicamente, el análisis realizado consiste en un procedimiento AMDF (*"Average Magnitude Difference Function"*) sobre la señal de voz filtrada paso-bajo a 800 Hz y posteriormente filtrada inversamente LPC con orden 2. Se consideran 60 posibles valores para frecuencia fundamental definidas en el rango de 50 Hz a 400 Hz de acuerdo con la tabla (5.7). Por lo tanto, en este caso nuestra fuente es discreta y toma valores en un conjunto finito de 60 valores. Para una secuencia de muestras $s(n)$, la función AMDF se define como

$$\mathrm{AMDF}(k) = \frac{1}{L}\sum_{j=1}^{L} |s(j) - s(j-k)| \qquad k = p_{\min}, \cdots, p_{\max} \tag{5.5}$$

donde para k se consideran los valores mostrados en la tabla (5.7), siendo $p_{\min} = 20$ y $p_{\max} = 156$. El periodo fundamental corresponderá al valor k que minimice la expresión anterior.

Igualmente, para la caracterización sonoro/no-sonoro, se ha utilizado el procedimiento considerado en el mismo standar, que tiene en cuenta los cruces por cero, la energía en la zona baja del espectro, coeficientes de reflexión, etc [Camp86]. Las estimaciones realizadas se suavizan posteriormente introduciendo un retardo adicional

k	f (Hz)	k	f (Hz)	k	f (Hz)	k	f (Hz)	k	f (Hz)	k	f (Hz)
20	400	30	266	40	200	60	133	80	100	120	67
21	381	31	258	42	190	62	129	84	95	124	65
22	364	32	250	44	184	64	125	88	91	128	63
23	348	33	242	46	174	66	121	92	87	132	61
24	333	34	235	48	167	68	118	96	83	136	59
25	320	35	228	50	160	70	114	100	80	140	57
26	308	36	222	52	154	72	111	104	77	144	56
27	296	37	216	54	148	74	108	108	74	148	54
28	286	38	210	56	143	76	105	112	71	152	53
29	276	39	205	58	138	78	103	116	69	156	51

Tabla 5.7: Posibles Valores del Periodo (k) y Frecuencia (f) Fundamentales Considerados.

de 2 tramas.

En la gráfica (5.9) se muestra el histograma o distribución de los periodos fundamentales para la secuencia de entrenamiento TIMIT-2, de los cuales el 61.4% correspondían a la categoría de sonoros. La secuencia de test utilizada (*core-test*) tiene una tasa de sonoridad del 62.5%.

5.3.1 Cuantización Lloyd-Max y Diferencial

De igual manera que para la secuencia de los α_N, en la gráfica (5.10) se presentan los resultados obtenidos para la fuente discreta de frecuencias fundamentales considerando un esquema Lloyd-Max y un esquema diferencial DPCM con orden de predicción 3 diseñado en bucle abierto. Estos dos sistemas sirven como referencia para la comparación de esquemas posteriores. Los resultados mostrados fueron resultado de considerar tan solo las tramas sonoras, es decir, el número de bits por muestra no incluye información acerca la clasificación sonoro/no-sonoro, sólo incluye información necesaria para cuantizar el periodo fundamental caso de ser una trama sonora.

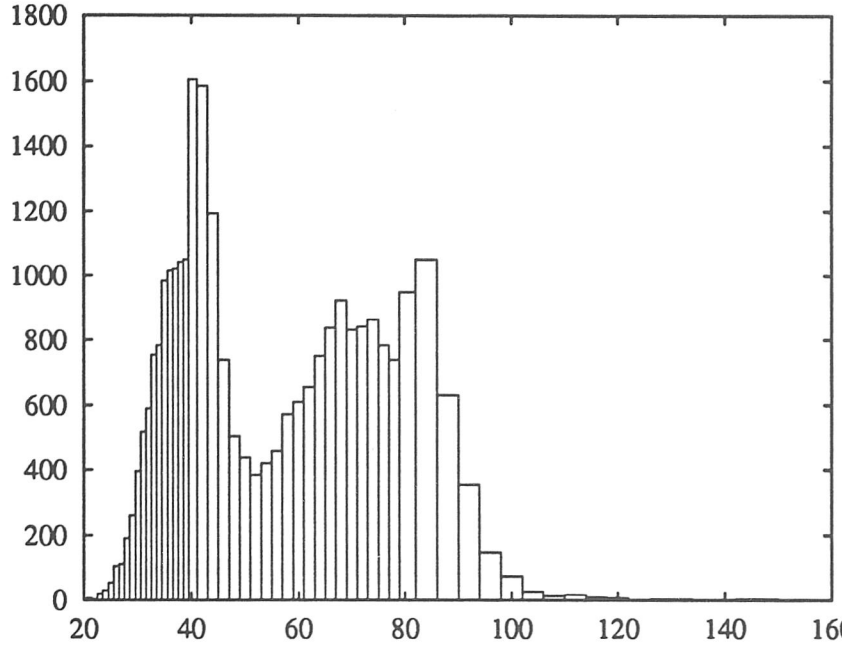

Figura 5.9: Histograma del Periodo Fundamental.

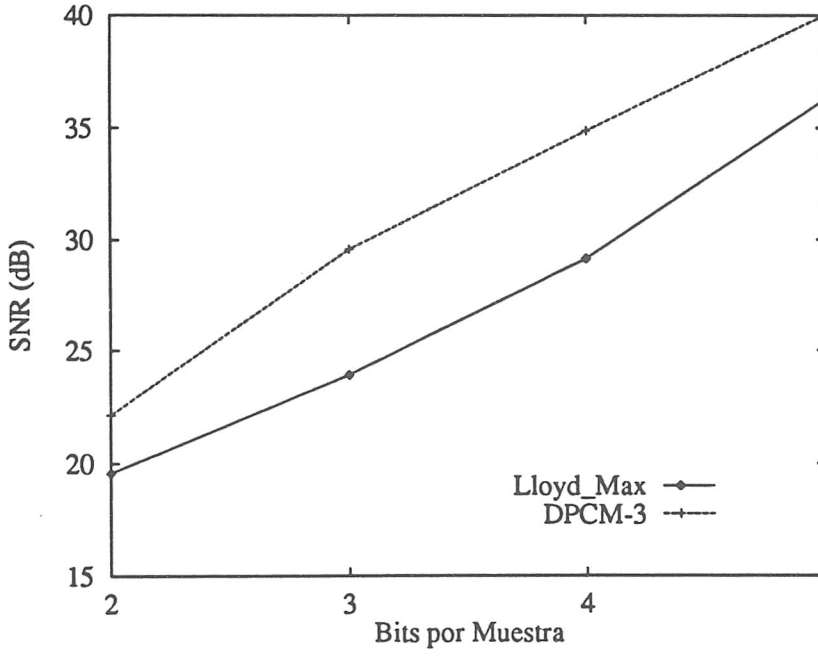

Figura 5.10: Cuantización DPCM y Lloyd-Max para el Periodo Fundamental.

En la tabla (5.8) se representan los coeficientes de predicción utilizados para dicho

ORDEN O	a_1	a_2	a_3
3	0.88	0.08	0.02

Tabla 5.8: Coeficientes de Predicción para el Periodo Fundamental.

esquema, obtenidos mediante el método de autocorrelación para la minimización de la energía residual.

5.3.2 Cuantización usando Codificación por "Trellis" (TCQ)

Por coherencia con el desarrollo realizado para la cuantización de la secuencia de α_n, en este subapartado se aplicará directamente el esquema denominado *TCQ predictivo con solapamiento* (ver apartado (5.2.4)). En la gráfica (5.11) se muestran los resultados obtenidos tras la aplicación de dicho esquema, para 4 estados con orden de predicción $O = 3$. En la decodificación se ha considerado una profundidad de 8 tramas o niveles y evidentemente como en casos anteriores se toma $S = 4$. Para este caso, las mejoras comparadas con el *DPCM-3* no son significativas.

5.3.3 Cuantización e Interpolación Combinadas

Al aplicar el algoritmo CQI escalar a la fuente generadora de periodos fundamentales se obtuvieron, para la secuencia de test considerada, unos resultados que no justifican la utilización directa del procedimiento CQI a la fuente considerada. La razón fundamental de este mal comportamiento estriba en que aun considerando umbrales U grandes, el algoritmo de dos pasadas tiene una tendencia natural a elegir como puntos ruptura muestras muy próximas, con lo que la longitud promedio de los bloques es pequeña. Recuérdese que dado un punto de ruptura para determinar el siguiente, tras estimar el siguiente del siguiente, se evalúan las posibles posiciones del primero y se elige la que minimiza la distorsión promedio de los dos bloques siguientes, considerando pertenecientes al bloque todas las muestras hasta el siguiente punto ruptura inclusive.

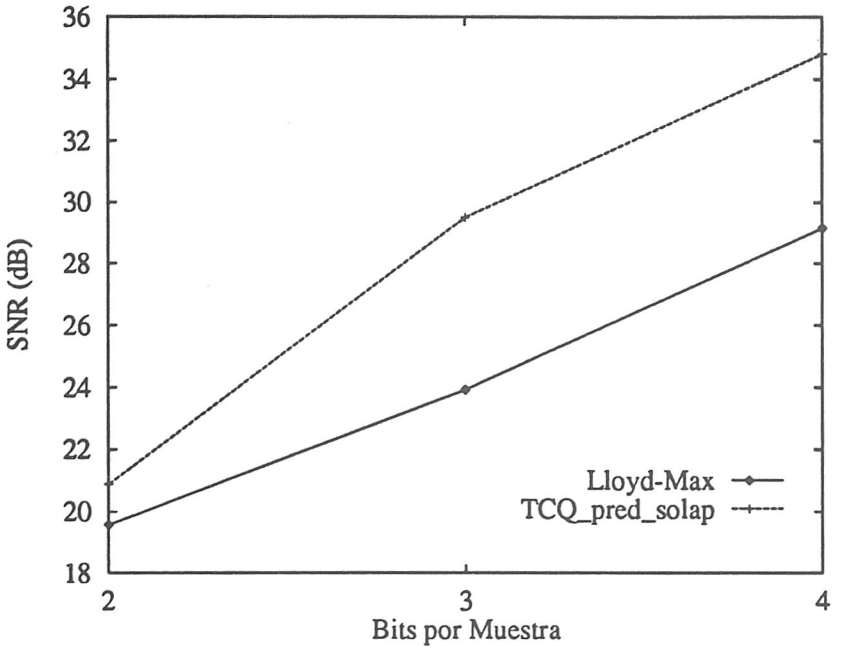

Figura 5.11: Cuantización TCQ Predictivo con Solapamiento y Lloyd-Max para el Periodo Fundamental.

Motivados por este mal comportamiento del algoritmo CQI para el periodo fundamental, se introduce la siguiente modificación consistente en una vez determinados los dos posibles puntos de ruptura siguientes, elegir como posición definitiva para el primero de ellos aquélla que estando más alejada del anterior implique una distorsión promedio, evaluada para todo el bloque, menor que el umbral U dado. La modificación introducida está reflejada en el nuevo diagrama de flujo de la figura (5.12) que utiliza la misma notación de la figura (4.16) para el caso vectorial.

En la gráfica (5.13) se muestran los resultados experimentales obtenidos con el nuevo esquema CQI, etiquetado NCQI, para la secuencia de periodos fundamentales. El número de bits por muestra para este esquema fué calculado teniendo en cuenta que las posiciones de los puntos ruptura se cuantizan uniformemente. Para la obtención de los puntos de la gráfica se consideraron distintos umbrales U, y para todos los casos se eligió un tamaño máximo para los bloques $N = 4$, que implica un retardo equivalente a la duración de 8 tramas.

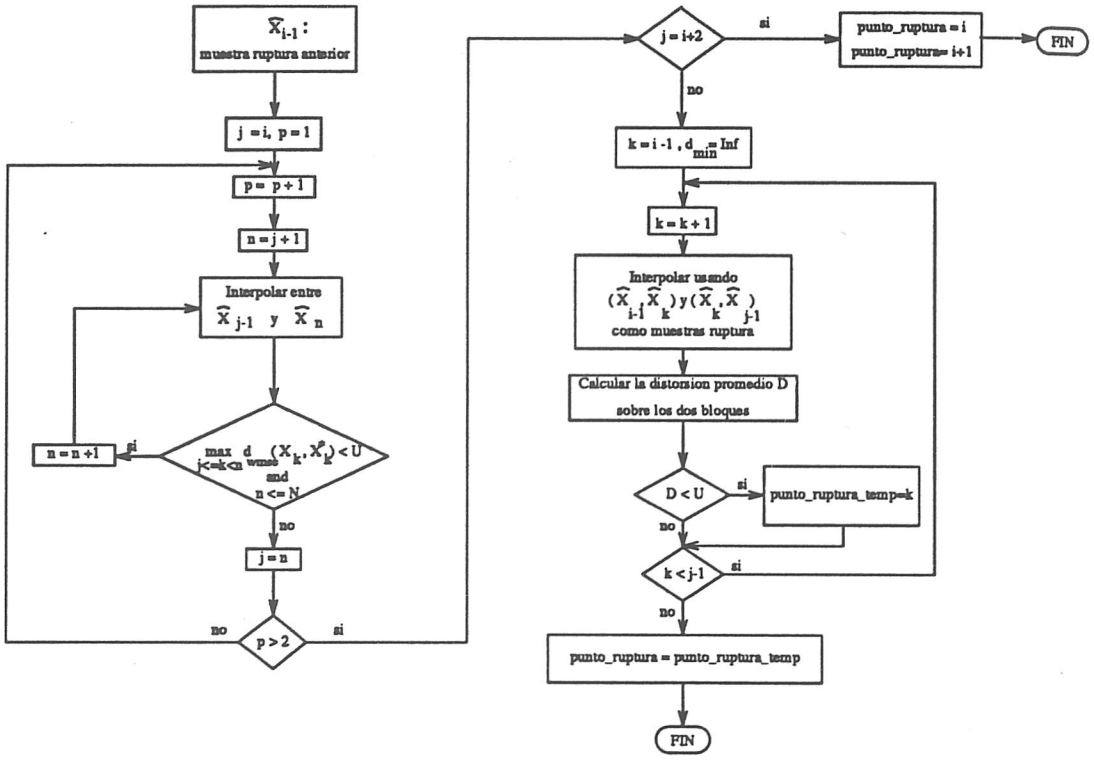

Figura 5.12: Diagrama de Flujo del Nuevo Algoritmo CQI.

Además de las anteriores modificaciones, se incorporaron las siguientes

- para suavizar las decisiones entre bloques adyacentes, una vez determinado el bloque definitivo, para la siguiente iteración del algoritmo en vez de considerar como muestra ruptura la última perteneciente al bloque en cuestión, es decir $\hat{X}_i = \hat{X}_{\text{punto\_ruptura}}$, se optó por promediar para todos los elementos del bloque anterior.

$$\hat{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_t \tag{5.6}$$

siendo T la longitud del bloque anterior y X_t las muestras después de la decodificación.

- Se consideró un diccionario de 6 bits, lo que significa considerar 64 niveles de cuantización, 4 más que los que la fuente discreta considerada presenta, por lo que para la codificación de las muestras ruptura el codificador introduce

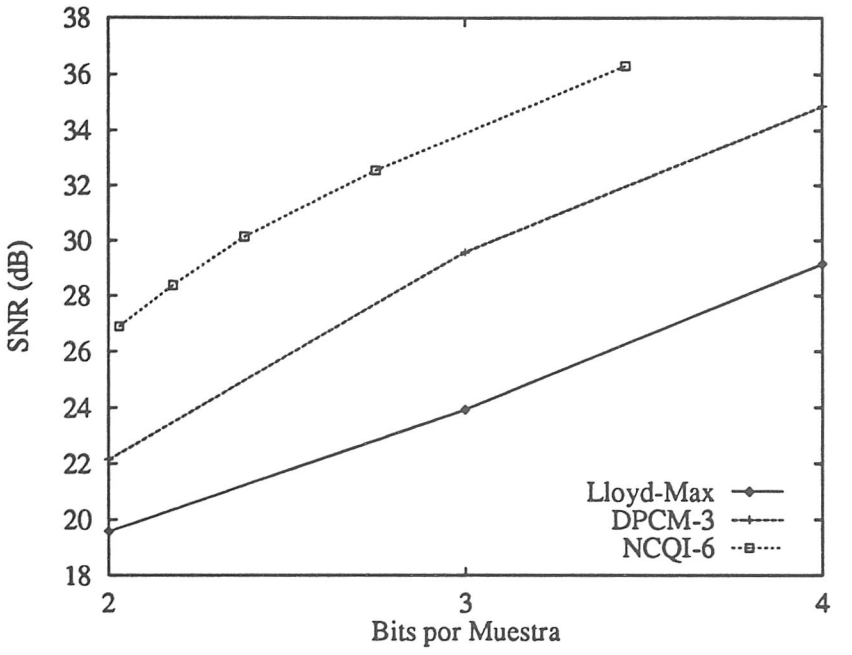

Figura 5.13: Algoritmo NCQI, DPCM-3 y Lloyd-Max para el Periodo Fundamental.

distorsión cero.

Por comparación, en la gráfica (5.13) se muestran también los resultados del esquema *DPCM-3* y el cuantizador Lloyd-Max.

Como resumen de los resultados obtenidos para el procedimiento NCQI y teniendo en cuenta la gráfica (5.13) se concluye que

- Con el procedimiento propuesto se consiguen mejoras de más de 4 dB respecto al esquema DPCM-3. Más de 6 dB se obtienen respecto al Lloyd-Max.
- En la gráfica no se ha incluido la información relativa a la caracterización sonoro/no-sonoro. Si se pretende transmitir dicha información sin error, es evidente que hay que considerar un bit más por muestra para incorporar dicha información.

Motivados por el ultimo punto, y para concluir con los resultados experimentales, en la gráfica (5.14) se muestran los resultados de considerar una modificación adicional al NCQI, (etiquetado *NCQIplus-6*) consistente en incorporar la caracterización

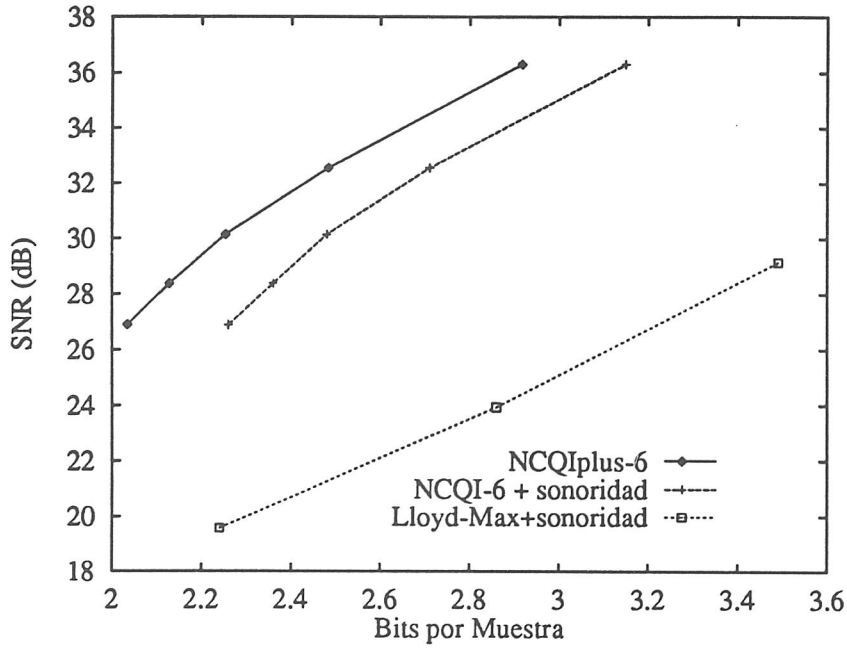

Figura 5.14: Algoritmo NCQIplus, NCQI Incluyendo Sonoridad y Lloyd-Max Incluyendo Sonoridad.

sonoro/no-sonoro dentro de los bits utilizados para transmitir el pitch. Recuérdese que se están considerando 6 bits y sólo 60 niveles de cuantización, luego una de las palabras no utilizadas se puede usar para transmitir dicha información. En dicha gráfica se incluyen los resultados de considerar el NCQI, pero incorporando un bit por trama para transmitir la sonoridad, que se etiqueta con *NCQI + sonoridad.* Así mismo se representan los resultados de incorporar dicha información usando un procedimiento Lloyd-Max (etiquetado *Lloyd-Max+sonoridad*). Finalmente, de la gráfica (5.14) se observa que al incorporar la información relativa a la sonoridad, para un umbral dado, se obtiene la misma SNR (como era de esperar) pero toda vez que la longitud media de los bloques aumenta, se obtiene un *bit-rate* promedio menor que con el esquema NCQI-6+sonoridad. Luego este procedimiento es apropiado para todo el rango de bits por muestra considerado comparado con transmitir un bit adicional por trama para caracterizar la sonoridad.

	CQI			CQI			VQ	Original
	LSP	α_N	pitch	LSP	α_N	pitch	solo	4 kHz
bit-rate (bps)	254.5	65.3	97.2	357.2	98.2	176.3	LSP	16 bits
bit-rate TOTAL	417.0			631.7			444.4	128 k
Inteligibilidad	31.0%			39.5%			37.3%	59.4%

Tabla 5.9: Test Subjetivo para el Esquema CQI y VQ.

5.4 Evaluación Subjetiva

Para concluir con este capítulo, en la tabla (5.9) se presentan los resultados obtenidos al realizar un test de inteligibilidad usando el mismo corpus y observadores que en los resultados mostrados en la tabla (4.12). En este caso se presentan los resultados obtenidos al aplicar el procedimiento CQI para codificar la información espectral, así como al codificar la excitación. Por comparación se incluyen las medidas de inteligibilidad obtenidas sobre el corpus original y la obtenida con el esquema VQ donde la excitación no es codificada.

Como se observa en la tabla (5.9), al codificar toda la información (excitación y espectro) con un *bit-rate* promedio de 631.7 bps, se obtiene una inteligibilidad superior a la obtenida al codificar tan solo los LSP, dejando la información correspondiente a la excitación sin codificar. Se puede concluir, por tanto, que la información correspondiente a la inteligibilidad está contenida principalmente en la información espectral. Recuérdese que utilizando el mismo umbral U, en la tabla (4.12), se obtuvo una inteligibilidad del 40.1% para el caso en el que solo se codifican los LSP.

La afirmación anterior se corrobora por los resultados obtenidos al codificar con un *bit-rate* total promedio de 417.0 bps. Obsérvese que la inteligibilidad obtenida es del 31.0%, que comparada con el 33.3% obtenido en la tabla (4.12) al codificar sólo los coeficientes LSP con el mismo umbral, supone una pequeña reducción en la inteligibilidad.

Capítulo 6

SUMARIO

Tras haber descrito y evaluado objetiva y subjetivamente los algoritmos propuestos junto a otros encontrados en la bibliografía, como conclusión en este último capítulo se resume toda la investigación realizada.

1. Para la señal de excitación se ha propuesto un modelo de análisis paramétrico en bucle cerrado que incorporando un criterio de fidelidad (como en los codificadores híbridos) adopta una parametrización igual a la considerada clásicamente en los *vocoders*. La mejor SSNR se ha obtenido con el llamado *análisis en bucle cerrado síncrono*, donde para las tramas sonoras la señal residuo está en fase con el tren de pulsos considerado.

2. Debido a que ninguna de las medidas objetivas puede sustituir completamente al proceso de percepción humano, para la evaluación objetiva de los codificadores de la información espectral se han utilizado la Distorsión Espectral, la Distancia de Itakura-Saito normalizada en ganancias y la Distancia Cepstral. Para la señal de excitación se consideró la Relación Señal Ruido.

3. Por sus propiedades en general y fundamentalmente por su facilidad para ser linealmente interpolados, para la información espectral se han elegido los coeficientes LSP como espacio de representación.

4. Tras las pruebas objetivas realizadas, la VQ con distancia IHM se ha mostrado eficaz para la reducción de las redundancias intra-trama en los coeficientes LSP. Para la fuente considerada, en el diseño de los diccionarios VQ se ha descartado la adopción de procedimientos de relajación estocástica debido a las insignificantes mejoras conseguidas en términos de evaluación objetiva.

5. Reduciendo las redundancias intra-trama mediante VQ Predictiva, se observa un mejor compromiso R-D que el obtenido mediante VQ. Esta mejora es tanto mayor cuanto mayor es el diccionario de predicción, lo que indica que cuanto mejor es la cuantización mayor es la ganancia de predicción. Razón esta por lo que la utilización de estos esquemas es más natural para regiones de mayor *bit-rate* a las aquí consideradas.

6. Se ha desarrollado un split-VQ predictivo conmutable con cuantización conmutable que comparado con un esquema split-VQ para regiones de mayor *bit-rate*

(22 bits por vector), obtiene unas mejoras del orden de 0.5 dB en la distorsión espectral al aumentar un bit por vector.

7. Siempre que la fuente a codificar lo permita, por extensión natural de los argumentos que justifican la VQ frente a la cuantización escalar, está justificado la consideración de una VQ generalizada que aproveche eficazmente las dependencias intra-trama. En particular para los LSP, por debajo de 300 bps, utilizando cuantización matricial se obtienen mejores resultados que la VQ-predictiva. Al aumentar el tamaño de las matrices código se observa una tendencia a la saturación, concluyéndose que pocas dependencias se pueden esperar para matrices de más de 3 tramas.

8. Se pueden obtener mejores resultados al incorporar un procedimiento de segmentación simultaneamente con la cuantización VQ generalizada, dando lugar a la cuantización por segmentos. Ahora bien, este tipo de aproximaciones están especialmente justificadas cuando el retardo máximo permisible no es una constricción prioritaria en el diseño y lo que es más, cuando la cuantización realizada es suficientemente aceptable, es decir cuando el tamaño de los diccionarios es elevado, lo cual introduce una complejidad adicional, pues en ese caso la secuencia de entrenamiento necesaria evidentemente es mucho mayor.

9. Para la cuantización VQ generalizada por segmentos es posible obtener mejores resultados si se realiza un alineamiento temporal dinámico que si se realiza estáticamente. Ahora bien, en este caso es necesario transmitir información lateral adicional para caracterizar el camino de alineamiento, lo cual hace no justificable su utilización.

10. Considerando tan solo las dependencias intra-trama lineales, en la región de 200 a 300 bps para los LSP, todavía se puede obtener un mejor compromiso R-D usando el esquema MF. Igualmente de entre todos los sistemas hasta ahora

mencionados, en la región de 300 a 500 bps para la información espectral el mejor compromiso se obtiene con el esquema MF-Split.

11. Al soslayar las limitaciones encontradas en el procedimiento MF con el esquema CQI , que combina de una forma eficaz los procedimientos de cuantización e interpolación, se consigue el mejor compromiso R-D en la región entre 200 y 400 bps para la información espectral.

12. Para la codificación de la información relacionada con la ganancia del filtro (parámetro α_N) se obtiene el mejor compromiso R-D (usando como medida de evaluación la SNR) particularizando el procedimiento CQI al caso monodimensional. Este algoritmo se ha comparado favorablemente con un procedimiento Lloyd-Max e incluso con un esquema de cuantización usando codificación por "trellis" predictiva con solapamiento.

13. Igualmente, para la codificación del periodo fundamental el mejor compromiso R-D se obtiene con una leve modificación del algoritmo CQI, como así se ha verificado tras las simulaciones realizadas. Adicionalmente, debido a que la clasificación de tramas en sonoras o no sonoras presenta un comportamiento estacionario, la incorporación de esta información dentro del nuevo algoritmo CQI se compara favorablemente con esquemas donde dicha caracterización se realiza mediante información lateral adicional.

14. Tras las pruebas de inteligibilidad realizadas, si bien la puntuación obtenida para el test de partida (sin codificar) es bastante baja debido a fenómenos de coarticulación, se ha constatado que la información correspondiente a la inteligibilidad está fundamentalmente contenida en la envolvente espectral.

Bibliografía

[Atal79] B.S. Atal and M.R. Schroeder. "Predictive Coding of Speech and Subjective Error Criteria,". *IEEE trans. on Acoustic, Speech and Signal Processing*, ASSP-27:247–254, June 1979.

[Atal82] B.S. Atal and J.R. Remde. "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates,". *IEEE ICASSP-82*, pages 614–617, April 1982.

[Atal83] B.S. Atal. "Efficient Coding of LPC Parameters by Temporal Decomposition,". *IEEE ICASSP-83, Boston*, pages 81–84, 1983.

[Atal85] B.S. Atal and M.R Schroeder. "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates,". *IEEE ICASSP-85*, pages 937–940, May 1985.

[Bar93] C.F. Barnes and R.L. Frost. "Vector Quantizers with Direct Sum Codebooks,". *IEEE Trans. on Information Theory*, 39(2):565–580, March 1993.

[Berger71] T. Berger. *Rate Distortion Theory, A Mathematical Basis for Data Compression.* Prentice-Hall, 1971.

[Berger73] T. Berger. "Rate Distortion Theory and Data Compression,". CISM Summer School Lecture Notes, 1973.

[Blaut72] R. E. Blaut. "Computation of Channel Capacity and Rate-Distortion Functions,". *IEEE Trans. on Information Theory*, IT-18(4):460–473, July 1972.

[Brands91] M. Brandstein, J. Hardwick, and J. Jim. *Advances in Speech Coding*, chapter The Multi-Band Excitation Speech Coder, pages 215–223. Kluwer Academic Publishers, 1991.

[Buzo80] A. Buzo, A.H. Gray, R.M. Gray, and J.D. Markel. "Speech Coding Based Upon Vector Quantization,". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-28(5):562–574, October 1980.

[CCITT90] UIT-Unión Internacional de Telecomunicaciones. "Modulación por impulsos Codificados Diferencial Adaptativa (MICDA) a 40, 32, 24, 16 kbit/s. Recomendación G.726.,", Diciembre 90.

[Camp86] J.P. Campbell and T. E. Tremain. "Voiced/Unvoiced Classification of Speech with Applications to the U.S. Goverment LPC-10E Algorithm,". In *IEEE ICASSP-86*, pages 473–476, 1986.

[Camp91] J.P. Campbell Jr., T.E. Tremain, and V.C. Welch. *Advances in Speech Coding*, chapter The DoD 4.8 KBPS Standard (Proposed Federal Standard 1016), pages 121–133. Kluwer Academic Publishers, 1991.

[Chang86] P-C. Chang and R.M. Gray. "Gradient Algorithms for Designing Predictive Vector Quantizers,". *IEEE Trans. on Acoustics, Speech, and Audio Processing*, ASSP-34(4):679–690, August 1986.

[Cheng91] Y.M. Cheng and D. O'Shaughnessy. "Short-Term Temporal Decomposition and its Properties for Speech Compression,". *IEEE Trans. on Signal Processing*, 39(6):1280–1290, June 1991.

[Cheng93] Y.M. Cheng and D. O'Shaughnessy. "On 450-600 b/s Natural Sounding Speech Coding,". *IEEE Trans. on Speech and Audio Processing*, 1(2):207–220, April 1993.

[Chou89] P.A. Chou, T. Lookabaugh, and R.M. Gray. "Optimal Prunning with Applications to Tree-Structured Source Coding and Modeling,". *IEEE Trans. on Information Theory*, 35:299–315, March 1989.

[Cuper85] V. Cuperman and A. Gersho. "Vector Predictive Coding of Speech at 16 kbits/s,". *IEEE Trans. on Communications*, COM-33(7):685–696, July 1985.

[Dimo89] S. Dimolitsas. "Objective Speech Distortion Measures and their Relevance to Speech Quality Assessments,". *IEE Proceedings I*, 136(5):317–324, October 1989.

[Dimo93] S. Dimolitsas. *Speech and Audio Coding for Wireless and Network Applications*, chapter Subjetive Assessment Methods for the Measurement of Digital Speech Coder Quality, pages 43–53. Kluwer Academic, 1993.

[Dun85] M.O. Dunham and R.M. Gray. "An Algorithm for the Design of Labeled-Transition Finite-State Vector Quantizers,". *IEEE Trans. on Communications*, COM-33(1):83–89, January 1985.

[FS1015] Federal Standard. "Telecommunications: Analog to Digital Conversion of Voice by 2400 Bit/Second Linear Predictive Coding,". Technical report, Department of Defense. US Goverment., 1984.

[Farvar89] N. Farvardin and R. Laroia. "Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transformation,". In *IEEE-ICASSP 1989*, 1989.

[Farvar90] N. Farvardin. "A Study of Vector Quantization for Noisy Channels,". *IEEE Trans. on Information Theory*, 36(4):779–809, July 1990.

[Farvar93] N. Farvardin. *Recent Advances and Trends in Speech Recognition and Coding*, chapter Speech Coding over Noisy Channels. Springer-Verlag, 1993. En prensa.

[Fono89] J.A. Rodríguez Fonollosa and E. Masgrau Gómez. *Cuantificación Vectorial Adaptativa Aplicada a la Codificación de Voz.* PhD thesis, Universidad Politécnica de Cataluña, Dept. de Teoría de la Señal y Comunicaciones, 1989.

[Forn73] G.D. Forney Jr. "The Viterbi Algorithm,". *Proceedings of the IEEE*, 61:268–278, March 1973.

[Furui81] S. Furui. "Cepstral Analysis Technique for Automatic Speaker Verification,". *IEEE Trans. on ASSP*, ASSP-29, 1981.

[Galvez91] J.A. Gálvez López and A.J. Rubio Ayuso. *Estimación del Pitch de Señales de Voz. Aplicación a la Identificación de Locutores.*, volume 19. Monografías del Dpto. de Electrónica y Tecnología de Computadores de la Universidad de Granada., Enero 1991.

[Gersho92] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.

[Gersho94] A. Gersho. "Advances in Speech and Audio Compression,". *Proceedings of the IEEE*, 82(6):900–918, June 1994.

[Golub83] G. H. Golub and C.F. Van Loan. *Matrix Computations.* North Oxford Academic, 1983.

[Gray80] R.M. Gray, A. Buzo, A.H. Gray Jr, and Y. Matsuyama. "Distortion Measures for Speech Processing,". *IEEE Trans. on ASSP*, ASSP-28(4):367–376, August 1980.

[Gray90] R.M. Gray. *Source Coding Theory.* Kluwer Academic Publishers, 1990.

[Honda92] M. Honda and Y. Shiraki. *Advances in Speech and Signal Processing*, chapter Very Low-Bit-Rate Speech Coding, pages 209–230. M. Dekker, Inc., 1992.

[Hussa93] Y. Hussain and N. Farvardin. "Variable-Rate Finite-State Vector Quantization and Applications to Speech and Image Coding,". *IEEE Trans. on Speech and Audio Processing*, 1(1):25–38, January 1993.

[Jaya84] N.S. Jayant and P. Noll. *Digital Coding of Waveforms (Principles and Applications to Speech and Video)*. Prentice-Hall, 1984.

[Jaya93] N. Jayant, J. Johnston, and R. Safranek. "Signal Compression Based on Models of Human Perception,". *Proceedings of the IEEE*, 81(10):1385–1422, October 1993.

[Juang82] B.-H. Juang, D.Y. Wong, and A.H. Gray Jr. "Distortion Performance of Vector Quantization for LPC Voice Coding,". *IEEE Trans. on ASSP*, ASSP-30(2):294–304, April 1982.

[Juang84] B.-H. Juang. "On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation,". *ATT Bell Labs. Technical Journal*, 63(8):1477–1498, October 1984.

[Kang82] G.S. Kang and S.S. Everett. "Improvement of Narrowband Linear Predictive Coder. Part 1- Analysis Improvements,". Technical Report NRL Report 8645, Naval Research Laboratory, December 1982.

[Kang84] G.S. Kang and S.S. Everett. "Improvement of Narrowband Linear Predictive Coder. Part 2- Synthesis Improvements,". Technical Report NRL Report 8799, Naval Research Laboratory, June 1984.

[Kang85] G.S. Kang and L.J. Fransen. "Low-Bit- Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs),". Technical Report NRL Report 8857, Naval Research Laboratory, January 1985.

[Kemp87] D.P. Kemp, J.P. Campbell, D.L. Andre, and D.J. Rahikka. "NSA LPC-10 Version 52,". Draft Documnentation, February 1987.

[Kemp91] D.P. Kemp, J.S. Collura, and T.E. Tremain. "Multi-Frame Coding of LPC Parameters at 600-800 bps,". *IEEE ICASSP-91*, pages 609–612, 1991.

[Kita82] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi. "Comparison for Objective Speech Qualyti Measures for Voiceband Codecs,". *IEEE ICASSP*, May 1982.

[Kita84] N. Kitawaki, M. Honda, and K. Itoh. "Speech-Qualyti Assesment Methods for Speech-Coding Systems,". *IEEE Communications Magazine*, 22(10):26–33, October 1984.

[Kita88] N. Kitawaki and H. Nagabuchi. "Qualyti Assessment of Speech Coding and Speech Synthesis Systems,". *IEEE Communications Magazine*, pages 36–44, October 1988.

[Kita91] N. Kitawaki. *Advances in Speech Signal Processing*, chapter Quality Assesment of Coded Speech, pages 357–385. M. Dekker Inc., 1991.

[Kroon91] P. Kroon and B.S. Atal. *Advances in Speech Coding*, chapter On Improving the Performance of Pitch Predictors in Speech Coding Systems, pages 321–327. Kluwer Academic Publishers, 1991.

[Kroon92] P. Kroon and B.S. Atal. *Advances in Speech and Signal Processing*, chapter Predictive Coding of Speech Using Analysis-by-Synthesis Techniques, pages 141–164. M. Dekker, Inc., 1992.

[Laroia91] R. Laroia, N. Phamdo, and N. Farvardin. "Robust and Efficient Quantization of Speech LSP Parameters using Structured Vector Quantizers,". In *IEEE ICASSP*, pages 641–644, 1991.

[Linde80] Y. Linde, A. Buzo, and R.M. Gray. "An Algorithm for Vector Quantizer Design,". *IEEE Trans. on Communications*, COM-28(1):84–95, January 1980.

[Lloyd82] S.P Lloyd. "Least Squares Quantization in PCM,". *IEEE Trans. on Information Theory*, IT-28(2):129–137, March 1982.

[Lopez90a] J.M. López Soler, A. Peinado Herreros, J.C. Segura Luna, V. Sanchez Calle, and A.J. Rubio Ayuso. "Bottom to Top Algorithms for Tree-Codebook Design in Vector Quantization for Speech Coding,". In Erdal Arikan, editor, *Communication Control and Signal Processing*, volume II, pages 1109–1113. Bilkent University, Elseiver, July 1990.

[Lopez90b] J.M. López Soler, A.M. Peinado Herreros, J.C. Segura Luna, V. Sánchez Calle, and A. Rubio Ayuso. "Distances Measures Performance in Vector Quantization,". In M.A. Lagunas L. Torres, E. Masgrau, editor, *Signal Processing V: Theories and Applications*, volume II, pages 1291–1294. Elseiver Science Publishers, September 1990.

[Lopez90c] J.M. López Soler and A.J. Rubio Ayuso. *Codificación de la Voz a Baja Velocidad de Transmisión*, volume 16. Monografías del Dpto. de Electrónica y Tecnología de Computadores de la Universidad de Granada., Enero 1990.

[Lopez93a] J.M. López Soler and N. Farvardin. "A Combined Quantization-Interpolation Scheme for Very Low Bit Rate Coding of Speech LSP Parameters,". In *IEEE ICASSP-93*, volume II, pages 21–24, April 1993.

[Makh75] J. Makhoul. "Linear Prediction: A Tutorial Review,". *Proceedings of the IEEE*, 63:561–580, April 1975.

[Makh85] J. Makhoul, S. Roucos, and H. Gish. "Vector Quantization in Speech Coding,". *Proceedings of th IEEE*, 73(11):1551–1588, November 1985.

[Marcell90] M.W. Marcellin and T.R. Fischer. "Trellis Coded Quantization of Memoryless and Gauss-Markov Sources,". *IEEE Transactions on Communications*, 38(1):82–93, January 1990.

[Mark72] J.D. Markel. "The SIFT Algorithm for Fundamental Frequency Estimation,". *IEEE Trans. Audio Electroacoustic*, AU-20:367–377, December 1972.

[Mark76] J.D. Markel and A.H. Gray Jr. *Linear Prediction of Speech.* Springer-Verlag, 1976.

[Marq90] J.S. Marques, L.B. Almeida, and J.M. Tribolet. "Harmonic Coding at 4.8 kb/s,". *IEEE ICASSP-90*, 1:17–20, April 1990.

[McAu92] R.J. McAulay. *Advances in Speech and Signal Processing*, chapter Low-Rate Speech Coding Based on the Sinusoidal Model, pages 165–208. M. Dekker, Inc., 1992.

[Myers81] C.S. Myers and L.R. Rabiner. "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition,". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-29(2):284–297, April 1981.

[Ney84] H. Ney. "The Use of a One-Stage Dynamic Progamming Algorithm for Connected Word Recognition,". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, assp-32(2):263–271, April 1984.

[Oppen75] A.V. Oppenheim and R.W. Shafer. *Digital Signal Processing.* Prentice-Hall, 1975.

[Pali91] K.K. Paliwal and B.S. Atal. "Efficient Vector Quantization of LPC Parameters at 24 bits/frame,". In *IEEE ICASSP-91*, pages 661–664, 1991.

[Panzer93] I.L. Panzer, A.D. Sharpley, and W.D. Voiers. *Speech and Audio Coding for Wireless and Network Applications*, chapter A Comparison of Subjective Methods for Evaluating Speech Quality, pages 59–65. Kluwer Academic, 1993.

[Peina94] A.M. Peinado Herreros, J.C. Segura Luna, and A.J. Rubio Ayuso. *Selección y Estimación de Parametros en Sistemas de Reconocimiento de Voz Basados en Modelos Ocultos de Markov.* PhD thesis, Universidad de Granada. Dept de Electrónica, Enero 1994.

[Peter90] P. Peterson, P. Jeanrenaud, and J. Vandergrift. "Improving Intelligibility of a 300 B/S Segment Vocoder,". *IEEE ICASSP-90*, pages 653–656, 1990.

[Pham90] N. Phamdo and N. Farvardin. "Coding of Speech LSP Parameters Using TSVQ with Interblock Noiseless Coding,". *IEEE ICASSP-90*, pages 193–196, 1990.

[Proak89] J.G. Proakis. *Digital Communications.* McGraw-Hill Book Company, second edition, 1989.

[Quack88] S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements. *Objetive Measures of Speech Quality.* Prentice-Hall, 1988.

[Rab78] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[Rab85] L.R. Rabiner and F.K. Song. "Single-Frame Vowel Recognition Using Vector Quantization with Several Distance Measures,". *ATT Bell Labs. Technical Journal*, 64(10):2319–2331, December 1985.

[Rose90] K. Rose, E. Gurewitz, and G. Fox. "A Deterministic Annealing Approach to Clustering,". *Pattern Recognition Letters*, 11(9):589–594, September 1990.

[Ross74] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley. "Average Magnitude Difference Function Pitch Extractor,". *IEEE Trans. on Acoustic, Speech and Signal Processing*, ASSP-22:353–362, October 1974.

[Rou82] S. Roucos, R. Schwartz, and J. Makhoul. "Vector Quantization for Very-Low-Rate Coding of Speech,". In *IEEE GlobeCom 82*, pages 1074–1078, December 1982.

[Rou83] S. Roucos, R.M. Schwartz, and J. Makhoul. "A Segment Vocoder at 150 b/s,". *IEEE ICASSP-83, Boston*, pages 61–64, 1983.

[Schroe85] M.R. Schroeder and B.S. Atal. "Code-Excited Linear Prediction (CELP) High-Quality Speech at Very Low Bit Rates,". In *IEEE ICASSP-85*, pages 937–940, March 1985.

[Segu90] J.C. Segura Luna, J. M. López Soler, A.M. Peinado Herreros, V. Sanchez Calle, and A.J. Rubio Ayuso. "Signal Segmentation into Spectral Homogeneous Units,". In L. Torres, E. Masgrau, and M.A. Lagunas, editors, *SIGNAL PROCESSING V: Theories and Appplications*, pages 1255–1258. Elseiver Science Publishers, 1990.

[Segu91] J.C. Segura Luna. *Modelos de Markov con Cuantización Dependiente para Reconocimiento de Voz.* PhD thesis, Universidad de Granada. Dept. de Electrónica, Noviembre 1991.

[Shira88] Y. Shiraki and M. Honda. "LPC Speech Coding Based on Variable-Length Segment Quantization,". *IEEE Trans. on Acoustics, Speech , and Signal Processing*, 36(9):1437–1444, September 1988.

[Shohan93] Y. Shohan. "High-Quality Speech Coding at 2.4 to 4.9 KBPS Based on Time-Frequency Interpolation,". In *IEEE ICASSP-93*, volume II, pages 167–170, 1993.

[Shore83] J.E.Shore and D.K. Burton. "Discrete Utterance Speech Recognition Without Time Alignment,". *IEEE Trans. on Information Theory*, IT-29(4):473, 1983.

[Soong84] F. K. Soong and B-H Juang. "Line Spectrum Pair (LSP) and Speech Data Compression,". In *IEEE ICASSP-84*, volume 1, pages 1.10.1–1.10.4, 1984.

[Soong88] F. K. Soong and B-H Juang. "Optimal Quantization of LSP Parameters,". In *IEEE ICASSP-88*, pages 394–397, 1988.

[Soong90] F. K. Soong and B-H Juang. "Optimal Quantization of LSP Parameters Using Delayed Decisions,". In *IEEE ICASSP-90*, volume 1, pages 185–188, 1990.

[Suga86] N. Sugamura and F. Itakura. "Speech Analysis and Synthesis Methods Developed at ECL in NTT -from LPC to LSP-,". *Speech Communication*, 5:199–215, June 1986.

[Suga88] N. Sugamura and N. Farvardin. "Quantizer Design in LSP Speech Analysis-Synthesis,". *IEEE Journal on Selected Areas in Communications*, 6(2):432–440, February 1988.

[T193] T1A1.6 Working Group on Specialized Signal Processing. "Technology-Independent, User-Oriented, Objective Assesment of Speech Transmission Quality,". Technical report, Exchange Carriers Standards Association, September 1993.

[Tran90] I.M. Trancoso, J.S. Marques, and C.M. Ribeiro. "CELP and Sinusoidal Coders: two Solutions for Speech Coding at 4.8-9.6 kbps,". *Speech Communication*, 9(5-6):389–400, December 1990.

[Trem82] T.E. Tremain. "The Goverment Standard Linear Predictive Coding Algorithm: LPC-10,". *Speech Technology*, pages 40–49, April 1982.

[Tsao85] C. Tsao and R.M. Gray. "Matrix Quantizer Design for LPC Speech Using the Generalized Lloyd Algorithm,". *IEEE Trans. on Acoustics, and Speech, and Signal Processing*, ASSP-33(3):537–545, June 1985.

[Tzeng90] F.F. Tzeng. "An Analysis-by-Synthesis Linear Predictive Model for Narrowband Speech Coding,". In *IEEE ICASSP-990*, pages 209–212, 1990.

[Tzeng91] F.F. Tzeng. *Advances in Speech Coding*, chapter Analysis-by-Synthesis Linear Predictive speech Coding at 4.8 Kbit/s and Below, pages 135–143. Kluwer Academic Publishers, 1991.

[Unger82] G. Ungerboeck. "Channel Coding with Multilevel/Phase Signals,". *IEEE Transactions on Information Theory*, IT-28(1):55–67, January 1982.

[Unger87a] G. Ungerboeck. "Trellis-Coded Modulation with Redundant Signal Sets Part I: Introduction,". *IEEE Communicactions Magazine*, 25(2):5–11, February 1987.

[Unger87b] G. Ungerboeck. "Trellis-Coded Modulation with Redundant Signal Sets Part II: State of the Art,". *IEEE Communicactions Magazine*, 25(2):12–22, February 1987.

[Vidal90] E. Vidal and A. Marzal. "A Review and new Approaches for Automatic Segmentation of Speech Signals,". In L. Torres, E. Masgrau, and M.A. Lagunas, editors, *SIGNAL PROCESSING V: Theories and Applications*, EUSIPCO-90, pages 43–53. Elseiver Science Publishers, September 1990.

[Viter79] A.J. Viterbi and J. K. Omura. *Principles of Digital Communication and Coding.* McGraw-Hill, 1979.

[Welch93] V.C. Welch and T.E. Tremain. "A New Goverment Standard 2400 bps Speech Coder,". In *IEEE Workshop on Speech Coding for Telecommu-*

nications. Speech Coding for the Network of the Future, pages 41–42, 1993.

[Wong82] D.Y. Wong, B-H Juang, and A.H. Gray. "An 800 bit/s Vector Quantization LPC Vocoder,". *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-30(5):770–780, October 1982.

[Yong88] M. Yong, G. Davidson, and A. Gersho. "Encoding of LPC Spectral Parameters Using Switched-Adaptive Interframe Vector Prediction,". *IEEE ICASSP-88*, pages 402–405, 1988.

[Zeger92] K. Zeger, J. Vaisey, and A. Gersho. "Globally Optimal Vector Quantizer Design by Stochastic Relaxation,". *IEEE Trans. on Signal Processing*, 40(2):310–322, February 1992.

[Zelins77] R. Zelinski and P. Noll. "Adaptive Transform Coding of Speech Signals,". *IEEE Trans. on ASSP*, pages 299–309, August 1977.