UNIVERSIDAD POLITECNICA DE CATALUÑA

Departamento de Teoria de la señal y comunicaciones

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

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

Barcelona, mayo 1993

Capítulo 3

TECNICAS ROBUSTAS DE REPRESENTACION DE LA SEÑAL DE VOZ

Tal como se ha visto en el capítulo anterior, el reconocimiento automático del habla requiere como primer paso la representación de la señal de voz mediante una sucesión de vectores de parámetros acústicos, equiespaciados en el eje temporal, con la información suficiente para poder identificar los sonidos en las siguientes etapas del sistema de reconocimiento. Esta etapa suele conocerse con el nombre de parametrización.

En la inmensa mayoría de sistemas de reconocimiento estos vectores de parámetros se obtienen a partir de un análisis espectral localizado de la señal de voz. Ello se debe a que tradicionalmente se ha supuesto que el oído humano es insensible a la fase de la transformada de Fourier localizada de la señal de voz [Sch75] y, por tanto, la información útil de la misma está contenida en su densidad espectral de potencia, que abreviadamente en esta memoria recibirá el nombre de espectro.

Debido a la inercia inherente a los órganos articulatorios es posible suponer que las características de la señal de voz no varían apreciablemente en un intervalo suficientemente corto de tiempo (del orden de 20 ms) y, por tanto, es posible realizar un análisis espectral cuasiestacionario sobre segmentos de señal de esta duración temporal. La evolución temporal de las características espectrales se obtiene repitiendo el análisis sobre segmentos consecutivos de la señal, que suelen tomarse con un cierto solapamiento temporal. De esta forma, a partir de la señal de voz se obtiene una secuencia de espectros, que pueden representarse mediante vectores.

El problema del análisis espectral, definido como la obtención de la distribución frecuencial de potencia de un proceso aleatorio a partir de ciertas medidas realizadas en un intervalo temporal finito de una de sus realizaciones, ha sido objeto de numerosos estudios en los últimos años, de los cuales han surgido infinidad de técnicas [Mar87]. Sin embargo, no es ni mucho menos un problema resuelto.

La predicción lineal de la señal de voz, basada en un modelado autorregresivo de la misma, ha mostrado gran utilidad en procesado de habla en general y, específicamente, en reconocimiento [Ita75]. Es la técnica de representación de señal de voz más utilizada en la actualidad debido a su correspondencia con el modelo de producción de la señal de voz y a su eficiencia y prestaciones.

Sin embargo, la técnica clásica de predicción lineal es muy sensible a la presencia de ruido aditivo y, por tanto, el comportamiento de los sistemas de reconocimiento cuya etapa de parametrización está basada en esta técnica se degrada rápidamente cuando el reconocimiento se realiza en condiciones ruidosas. Por ello, es necesario buscar nuevas técnicas más robustas de análisis espectral de la señal de voz.

En este capítulo se presentará la predicción lineal de la parte causal de la secuencia de autocorrelación de la señal de voz como una técnica de parametrización robusta del habla en presencia de ruido, estrechamente relacionada con la técnica de Coherencia Modificada Localizada (SMC, *Short-Time Modified Coherence*), propuesta por Mansour y Juang [Man89a], y con el uso de un sistema sobredeterminado de ecuaciones de Yule-Walker [Cad82]. Su uso en reconocimiento de habla ruidosa es muy interesante debido a su simplicidad, su eficiencia computacional y sus altas tasas de acierto, como se verá en los resultados experimentales presentados en el capítulo 6 de esta memoria.

Además de esta aproximación al problema, consistente en realizar un análisis espectral robusto de la señal de voz desde el punto de vista del procesado de la señal, otra forma de obtener parametrizaciones robustas de la señal de voz consiste en emular la capacidad auditiva humana, basándose en el hecho bien conocido de que nuestro oído parece percibir la voz mejor que cualquier máquina en presencia de ruido interferente sin un conocimiento previo de la voz ni del ruido. Dentro de este enfoque, una posibilidad es realizar una transformación de la escala de frecuencias que aproxime la sensibilidad logarítmica en frecuencia del oído, lo cual puede realizarse eficientemente mediante una transformación bilineal en el plano de frecuencias complejas. En este

trabajo, se estudiará el comportamiento de la transformación bilineal de frecuencias en reconocimiento de habla ruidosa.

El contenido de este capítulo está estructurado del siguiente modo. En el apartado 3.1 se revisan los modelos digitales de producción de la señal de voz, basados en los principios fisiológicos y en las características temporales y frecuenciales de la misma. El apartado 3.2 se resumen los principales características de la predicción lineal clásica. El apartado 3.3 está dedicado al tema de la sensibilidad al ruido de las técnicas de predicción lineal clásicas y las principales variaciones que se han propuesto para combatir el problema. En el apartado 3.4 se expone una nueva interpretación de las técnicas anteriores desde el punto de vista de la señal de autocorrelación, que dará pie a la introducción en el apartado 3.5 de la técnica de predicción lineal de la parte causal de la autocorrelación como parametrización robusta de la señal de voz en presencia de ruido. Finalmente, en el apartado 3.6 se aborda el tema de la transformación de la escala de frecuencias.

3.1.MODELADO DIGITAL DE PRODUCCION DE LA SEÑAL DE VOZ

En este apartado se revisan los principios fisiológicos básicos de producción del habla, las características temporales y frecuenciales de la señal de voz y el modelo digital de producción de la voz basado en los mismos, que sirve de fundamento a la aplicación a las técnicas de predicción lineal a la parametrización de la señal de voz.

3.1.1. PRINCIPIOS FISIOLOGICOS BASICOS

La voz es una onda acústica de presión que se origina a partir de los movimientos fisiológicos voluntarios de los órganos del aparato fonador humano. En todo tipo de sonidos, el aire es expelido desde los pulmones a la tráquea y forzado a pasar entre las cuerdas vocales. A partir de este momento, el estado de relajación o tensión de las cuerdas vocales y el movimiento relativo de los órganos articulatorios define los diferentes sonidos.

Durante la generación de los sonidos sonoros, el aire expelido hacia los labios por los pulmones provoca la vibración de las cuerdas vocales a un ritmo que depende de la presión del aire en la tráquea y del ajuste fisiológico de las mismas. Este ajuste incluye cambios en la longitud, grosor y tensión de las cuerdas vocales. El ritmo a que

se abre y cierra la glotis, orificio que queda entre las cuerdas vocales, se corresponde con la frecuencia fundamental de la voz, inversa del período observado en la señal acústica, y con el tono percibido (*pitch*, en la literatura inglesa). La presión del aire subglótica y las variaciones temporales del área glotal determinan la velocidad volumétrica del flujo de aire glotal expelida al tracto vocal. Esta velocidad volumétrica glotal define la entrada de energía acústica o función de excitación al tracto vocal.

El tracto vocal, que se extiende desde la glotis hasta los labios, actúa como un tubo acústico de sección no uniforme y variante con el tiempo. Esta variación temporal de la forma del tracto vocal es debida a los movimientos de los labios, la mandíbula, la lengua y el velo. Durante la generación de los sonidos no nasales, el velo separa el tracto vocal de la cavidad nasal. La cavidad nasal constituye un tubo acústico adicional para la transmisión del sonido usado en la generación de los sonidos nasales.

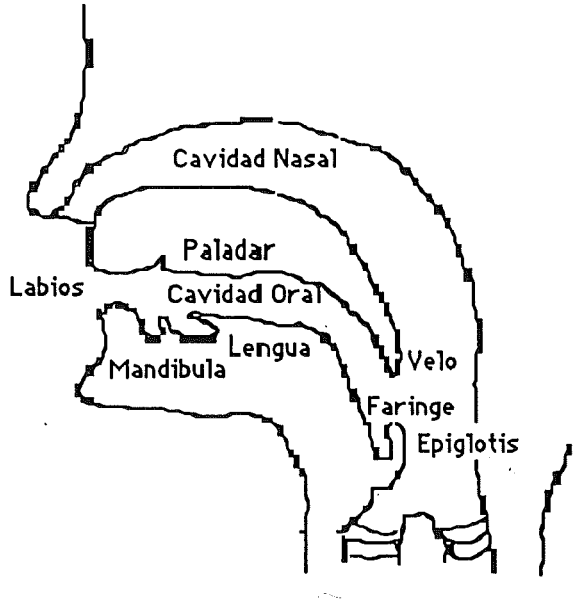


Fig. 3.1. Esquema del aparato fonador humano

Los sonidos sordos se generan manteniendo abiertas las cuerdas vocales voluntariamente, haciendo pasar el aire a través de ellas y usando los órganos articulatorios para crear una constricción. En la generación de los sonidos sonoros fricativos se produce a la vez vibración de las cuerdas vocales y constricción. Por

último, los sonidos oclusivos son generados provocando presión en la boca y liberando luego el aire abruptamente.

3.1.2. LA SEÑAL DE VOZ

Para ilustrar las implicaciones acústicas del proceso de producción de voz en los dominios temporal y frecuencial, en la figura 3.2 se ha representado la evolución temporal y frecuencial de la semisílaba /o/-/s/ de la palabra "dos" pronunciada en catalán. Para ello, la señal de voz fue filtrada de 100 a 3400 Hz con un filtro *antialiasing*, muestreada a 8 kHz y cuantificada con dos bytes.

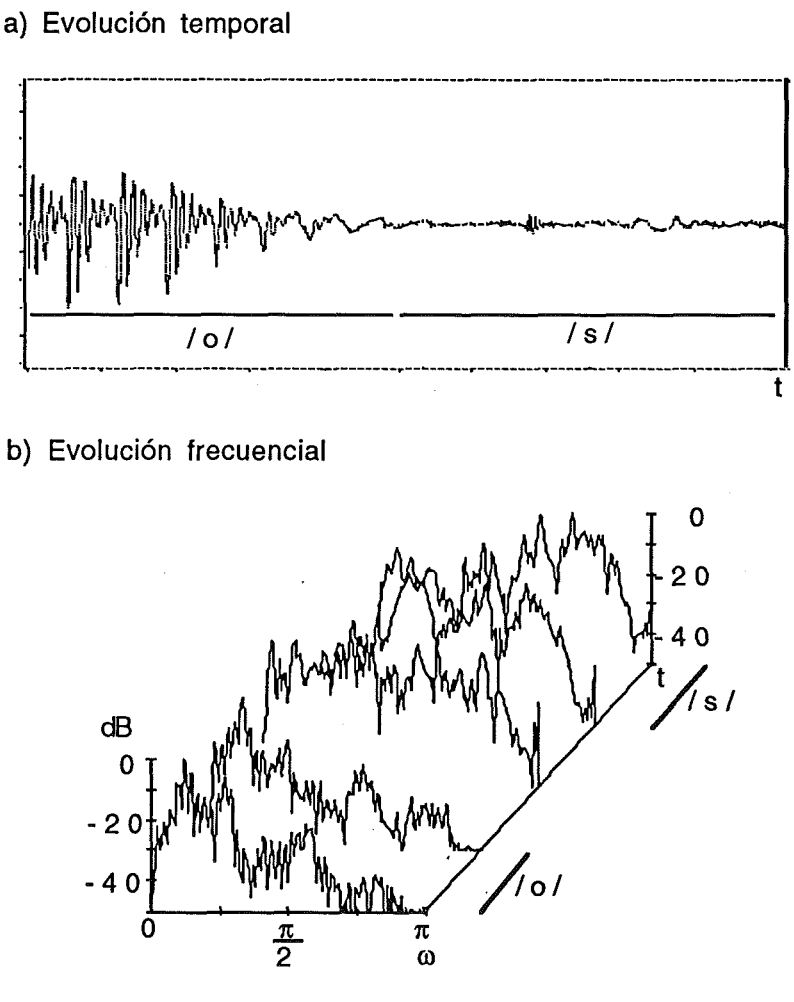


Fig. 3.2. Representación de la evolución temporal (a) y frecuencial (b) de la semisílaba /o/-/s/ extraída de la palabra "dos" en catalán.

La figura 3.2.a) representa la evolución temporal de las muestras de la señal de voz. Puede observarse que la parte estacionaria del sonido sonoro /o/ es aproximadamente periódica. La distancia entre los picos mayores muestra el periodo P de las vibraciones glotales. La frecuencia de las oscilaciones decrecientes de cada período determina la localización aproximada de la resonancia más importante del tracto vocal en el dominio frecuencial. Por otro lado, la señal correspondiente al sonido sordo /s/ no exhibe ninguna periodicidad, ya que en su generación no se produce vibración de las cuerdas vocales.

En la figura 3.2.b) están representados los logaritmos de los espectros (periodogramas) de cinco segmentos equiespaciados de la señal de voz anterior. La duración de estos segmentos es de 30 ms, lo cual permite suponer estacionariedad local. Puede observarse que los espectros correspondientes a segmentos de voz sonora presentan un detalle fino consistente en armónicos cada 1/P unidades de frecuencia que son debidos a la periodicidad mencionada. En cambio, en el caso de voz sorda el detalle fino del espectro tiene un cariz errático o ruidoso por la ausencia de periodicidad.

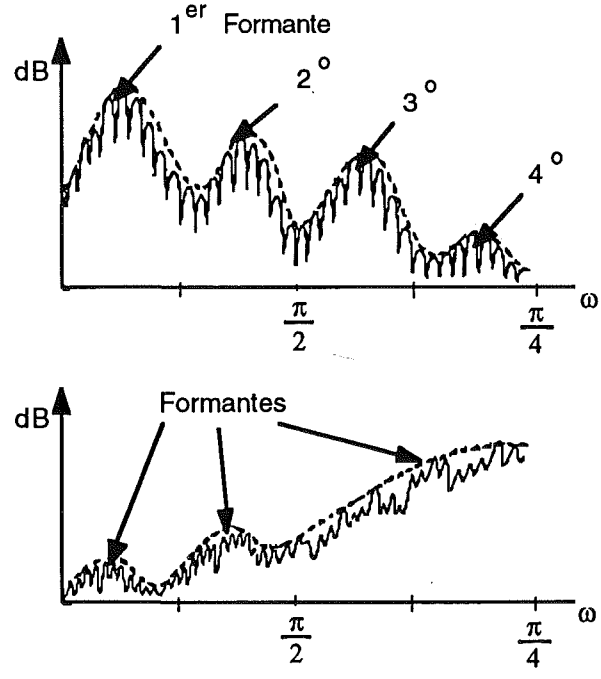


Fig. 3.3. Espectros típicos de voz sonora (arriba) y sorda (abajo)

En la figura 3.3. se representan los espectros típicos de un segmento de voz sonora (arriba) y sorda (abajo). Además de las estructuras finas características de cada tipo de espectro, se ha dibujado una envolvente suave superpuesta. Los picos de esta envolvente se denominan formantes y, esencialmente, se corresponden con las resonancias del tracto vocal.

A partir de estas figuras, se concluye que la señal de voz tiene una estructura compleja. Para modelar esta estructura, serían deseables modelos lineales e invariantes con el tiempo. Desafortunadamente, el mecanismo del habla no satisface ninguna de estas dos propiedades. El habla es un proceso que varía continuamente con el tiempo. Además, la glotis está acoplada al tracto vocal, lo cual da lugar a características no lineales. Sin embargo, haciendo algunas suposiciones razonables, es posible desarrollar modelos lineales invariantes con el tiempo sobre cortos intervalos de tiempo.

El modelo de producción del habla que se describirá en el siguiente apartado separa la estructura fina del espectro de su envolvente y asigna a cada componente del modelo un significado fisiológico. También se verá que esta envolvente puede obtenerse eficientemente mediante la predicción lineal de la señal de voz.

3.1.3. MODELO LINEAL DE PRODUCCION DEL HABLA

A finales de los años 50, Fant desarrolló un modelo lineal de producción del habla, que se representa esquemáticamente en la figura 3.4. Los supuestos en que se basa este modelo se exponen en detalle en [Fan60] y [Fla72].

La señal de velocidad volumétrica glotal $u_G(t)$ se modela como la salida de un filtro paso-bajo de dos polos con una frecuencia de corte de unos 100 Hz. La entrada a este filtro u(t) es un tren de impulsos de período P para sonidos sonoros y ruido aleatorio de espectro plano para el caso de sonidos sordos. No se considera la mezcla de excitaciones necesaria para la producción de los sonidos fricativos sonoros.

El tracto vocal se modela como un sistema todo-polos formado por una cascada de un pequeño número de resonadores de dos polos. Cada resonancia se define como un formante con su frecuencia central y su ancho de banda correspondientes. No se considera el efecto de la cavidad nasal en la producción de los sonidos nasales.

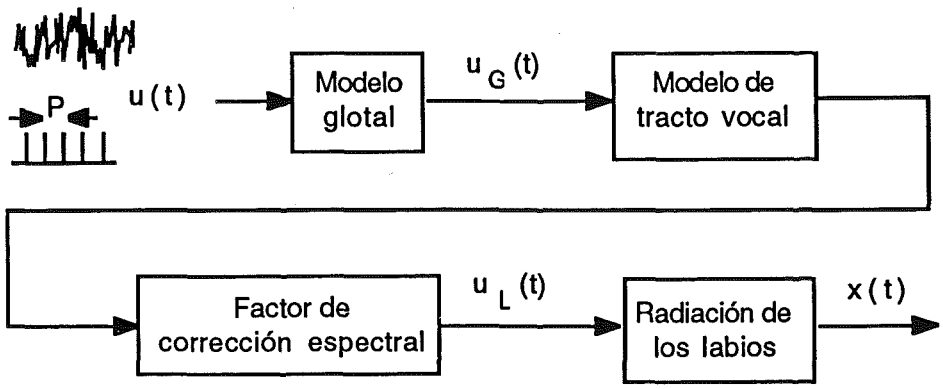


Fig. 3.4. Modelo lineal de producción del habla de Flant

Un modelado más preciso requiere un número infinito adicional de resonancias cuyo principal efecto en las frecuencias más bajas es elevar el nivel del espectro. Cuando sólo se modela de forma precisa el comportamiento a bajas frecuencias del sistema, por ejemplo, la banda de audio de 20 Hz a varios kHz, este efecto puede tenerse en cuenta mediante un factor de corrección espectral.

La señal de velocidad volumétrica en los labios $u_L(t)$ se transforma en una señal de presión acústica x(t) a una cierta distancia de los labios, a través del modelo de radiación de los labios.

Suponiendo invarianza con el tiempo, este modelo puede describirse en notación de transformada Z para su implementación discreta mediante la siguiente ecuación

$$X(z) = U(z)\ G(z)\ V(z)\ L(z), \qquad (3.1)$$

donde X(z) y U(z) son las transformadas Z de las secuencias discretas x(n) y u(n), resultantes de muestrear x(t) y u(t) a un período de muestreo T, y G(z), V(z) y L(z) son las funciones de transferencia de los sistemas discretos que modelan los efectos de la glotis, el tracto vocal y los labios, respectivamente. Hay que hacer notar que en la representación discreta puede eliminarse el factor de corrección espectral que figuraba en el modelo original [Rab68].

Una importante simplificación de este modelo consiste en combinar los efectos de la glotis, el tracto vocal y los labios y representarlos mediante una única función de transferencia H(z), es decir,

$$X(z) = U(z)\ H(z) \tag{3.2}$$

En la práctica, en la mayoría de las aplicaciones se modela el filtro H(z) como un filtro todo-polos

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}}. \tag{3.3}$$

La razones fundamentales por las que se utiliza un modelado todo-polos son:

a) Si se ignoran los sonidos nasales y algunos fricativos, la función de transferencia del tracto vocal es una función todo-polos y el efecto de la glotis y la radiación de los labios puede caracterizarse mediantes algunos polos adicionales.

b) Los parámetros de un modelo todo-polos pueden obtenerse eficientemente aplicando técnicas de predicción, a las que se dedicarán los siguientes apartados. Sin embargo, la utilización de modelos con ceros finitos conlleva la resolución de sistemas de ecuaciones no lineales, lo cual incrementa considerablemente el coste de cálculo.

c) Un modelo todo-polos permite aproximar cualquier modelo racional utilizando un número suficientemente elevado de polos.

A pesar del carácter no estacionario de la señal de la señal de voz, la aplicación de este modelo es posible gracias a la inercia inherente a los órganos articulatorios, la cual permite suponer que las características de la señal de voz no varían apreciablemente en un intervalo suficientemente corto de tiempo (del orden de 20 ms). Por ello, todos los parámetros del modelo son actualizados periódicamente.

El modelo de producción de voz simplificado está representado en la figura 3.5. El sistema es excitado por un tren de impulsos en el caso de voz sonora o por ruido en el caso de voz sorda. Los parámetros del modelo son la decisión sordo/sonoro, el tono en su caso y la ganancia G y los coeficientes $\{a_k\}_{k=1...p}$ del filtro H(z).

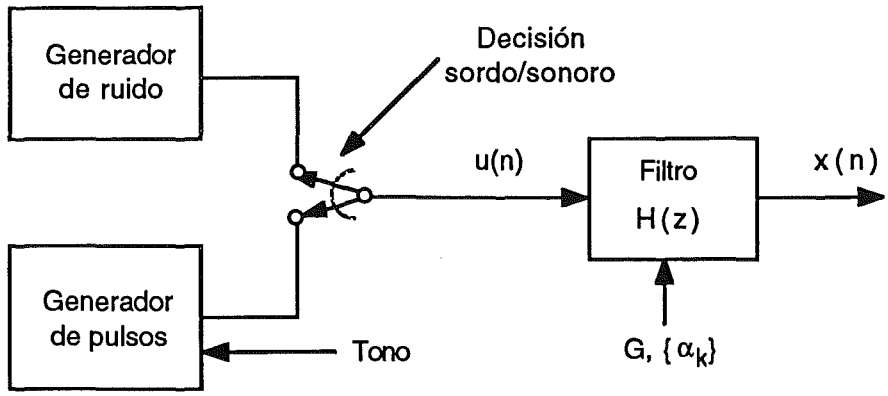


Fig. 3.5. Modelo simplificado de producción del habla

A partir de la expresión (3.2), el espectro de la señal de voz x(n) puede escribirse como

$$S_{xx}(\omega) = S_{uu}(\omega)\ |H(e^{j\omega})|^2, \qquad (3.4)$$

donde $S_{uu}(\omega)$ es el espectro de la excitación u(n) y $H(e^{j\omega})$ es la respuesta frecuencial del filtro.

Aplicando las técnicas de predicción lineal sobre la señal de voz se extraen de una manera simple los parámetros del filtro H(z), como se verá en los apartados siguientes. De este modo, se pueden separar del espectro $S_{xx}(\omega)$ eficientemente la estructura fina y la envolvente (ver figura 3.3), determinados por $S_{uu}(\omega)$ y $|H(e^{j\omega})|^2$, respectivamente. En el dominio temporal, esto equivale a deconvolucionar la señal de voz x(n), es decir, separar la excitación u(n) y la respuesta impulsional del filtro H(z), h(n), relacionados por la siguiente ecuación de convolución

$$x(n) = u(n) * h(n); \qquad (3.5)$$

o, en otras palabras, separar la información de sonoridad y tono de la estructura de formantes. Este hecho es de gran interés en reconocimiento del habla, ya que usualmente se utilizan vectores de parámetros acústicos relacionados con la envolvente espectral de la señal de voz.

Teniendo en cuenta que los parámetros del modelo son reestimados y actualizados periódicamente debido a la no estacionariedad de voz, una señal de voz se corresponde con una sucesión de vectores de parámetros acústicos equiespaciados en el eje temporal. La aplicación de este modelo justifica, pues, la parametrización de la señal de voz mediante técnicas de predicción lineal.

Las simplificaciones que conducen a este modelo conllevan lógicamente una serie de limitaciones. En primer lugar, está la cuestión de la variación de los parámetros. En sonidos continuos, como las vocales, los parámetros cambian muy lentamente. Sin embargo, en sonidos transitorios, como los oclusivos, el modelo no es tan bueno pero todavía adecuado. Una segunda limitación es la falta de ceros, que teóricamente se requieren para las nasales y algunas fricativas. En tercer lugar, la simple dicotomía de excitación sorda/sonora no es adecuada para sonidos sonoros fricativos. Otro problema es el desacoplo supuesto entre fuente y filtro. Afortunadamente, ninguna de estas deficiencias del modelo limita su aplicabilidad en la gran mayoría de los casos.

3.2. PREDICCION LINEAL CLASICA DE LA SEÑAL DE VOZ

Desde que el término predicción lineal fue acuñado por Wiener, esta técnica ha sido profusamente empleada en una amplio rango de aplicaciones bajo distintas formulaciones. Utilizada por primera vez para el análisis y síntesis del habla por Saito e Itakura [Sai66] y Atal y Schroeder [Ata67], ha producido un gran impacto en todos los aspectos del tratamiento del habla [Mar76].

La técnica de predicción lineal, abreviadamente LPC (*Linear Predictive Coding*), consiste en estimar el valor actual de una señal x(n) como una combinación lineal de las muestras anteriores. El valor estimado $\hat{x}(n)$ se escribe como

$$\hat{x}(n) = -\sum_{k=1}^{p} a_k x(n-k), \qquad (3.6)$$

donde p es el orden de predicción y a_k son los coeficientes de predicción. El problema básico de la predicción lineal consiste en determinar estos coeficientes a_k de forma que la aproximación de x(n) sea suficientemente buena de acuerdo con algún criterio.

El error entre la valor real x(n) y el valor estimado $\hat{x}(n)$ se denomina error de predicción y viene dado por la expresión

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^{p} a_k x(n-k). \qquad (3.7)$$

A partir de esta expresión, puede considerarse el error de predicción como respuesta a x(n) de un sistema, que se denomina filtro de error de predicción, cuya función de transferencia es

$$A(z) = 1 + \sum_{k=1}^{p} \alpha_k z^{-k} \qquad (3.8)$$

Además, a partir de (3.7) también puede escribirse

$$x(n) = - \sum_{k=1}^{p} a_k x(n-k) + e(n). \qquad (3.9)$$

Por tanto, el modelo de predicción lineal de generación de señal puede representarse como

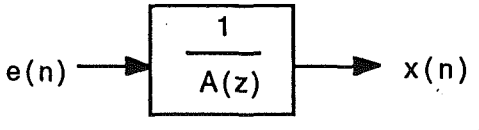


Fig. 3.6. Modelo de generación de señal de predicción lineal

Por otro lado, si la señal de voz obedece al modelo simplificado de la figura 3.5, combinando las expresiones (3.2) y (3.3) del apartado anterior se obtiene

$$X(z) = \frac{G\,U(z)}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}}. \qquad (3.10)$$

Tomando transformada Z inversa a ambos lados de la igualdad, puede escribirse (3.10) como

$$x(n) = -\sum_{k=1}^{p} \alpha_k x(n-k) + G\ u(n). \qquad (3.11)$$

Comparando las expresiones (3.9) y (3.11), se obtiene que si la señal de voz obedece al modelo mencionado y $\alpha_k = a_k$, entonces $e(n) = G\ u(n)$. Por tanto, el filtro de error de predicción A(z) será un filtro inverso del filtro H(z) de la expresión (3.3), es decir

$$H(z) = \frac{G}{A(z)} \qquad (3.12)$$

El problema básico de la predicción lineal de la señal de voz es la determinación del conjunto de coeficientes a_k directamente de la señal de tal forma que se obtenga una buena estimación de las propiedades espectrales de la señal de voz mediante el uso de (3.12).

Debido a la no estacionariedad de la señal de voz, si se descartan los métodos de estimación secuenciales, los coeficientes de predicción deben ser estimados sobre segmentos cortos de señal de voz, que se denominarán tramas. La estimación de máxima verosimilitud es difícil de obtener, por lo cual han surgido una gran variedad de formulaciones alternativas. La más común es la estimación de mínimos cuadrados. Fundamentalmente, se pueden distinguir dos tipos de estimadores de mínimos cuadrados: los que utilizan exclusivamente la predicción lineal hacia adelante (*forward*, en la literatura inglesa), que es la presentada en este apartado, y los que combinan esta con la predicción lineal hacia atrás (*backward*), que es análoga a la anterior pero considerando las p muestras futuras en lugar de las pasadas.

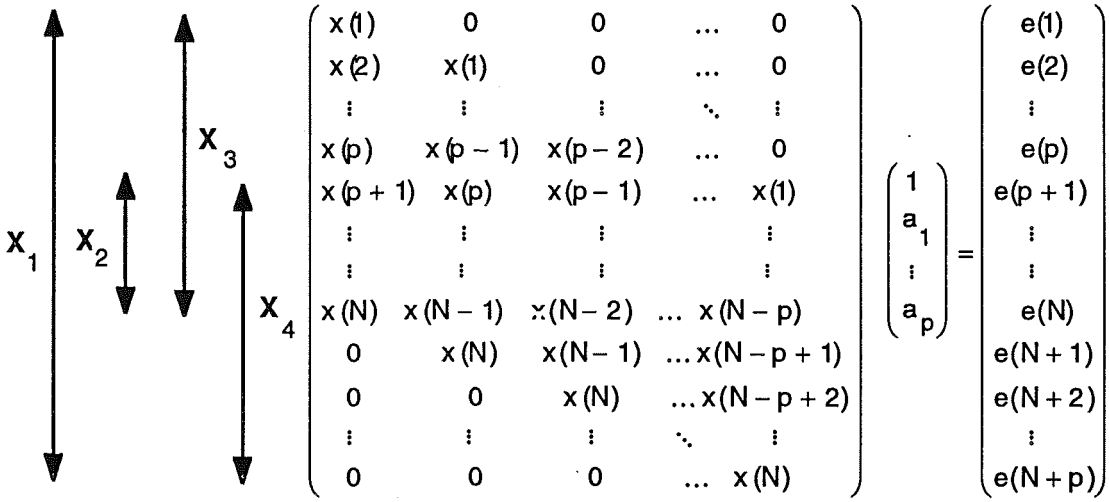
En el apartado 3.21 se revisarán los estimadores del primer tipo, especialmente los métodos de autocorrelación y covarianza. No se abordará el cálculo de la ganancia G del filtro H(z), ya que este parámetro no es usado en reconocimiento. Tampoco se revisarán los estimadores del segundo tipo, entre los que pueden destacarse el método de covarianza modificada y el método de Burg [Bur67], debido a que para longitudes de trama del orden de las usadas en reconocimiento del habla las prestaciones son muy similares a los métodos anteriores, más simples (ver capítulo 6).

Finalmente, el apartado 3.2.2 se dedicará a las propiedades del modelado espectral de la predicción lineal.

3.2.1. ESTIMACION DE MINIMOS CUADRADOS

Si se dispone de una trama de muestras de señal de longitud N, suponiendo x(n)=0 para n < 1 y n > N, se puede calcular e(n) a partir de la expresión (3.7) desde n = 1 hasta n = N+p. Fuera de este rango el error e(n) es nulo.

Utilizando formulación matricial, podemos escribir estos cálculos como



$$\begin{pmatrix} x(1) & 0 & 0 & \dots & 0 \\ x(2) & x(1) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x(p) & x(p-1) & x(p-2) & \dots & 0 \\ x(p+1) & x(p) & x(p-1) & \dots & x(1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ x(N) & x(N-1) & x(N-2) & \dots & x(N-p) \\ 0 & x(N) & x(N-1) & \dots & x(N-p+1) \\ 0 & 0 & x(N) & \dots & x(N-p+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & x(N) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} e(1) \\ e(2) \\ \vdots \\ e(p) \\ e(p+1) \\ \vdots \\ \vdots \\ e(N) \\ e(N+1) \\ e(N+2) \\ \vdots \\ e(N+p) \end{pmatrix} \qquad (3.13)$$

Abreviadamente,

$$\mathbf{X}\,\mathbf{A} = \mathbf{E}, \qquad (3.14)$$

donde **X** es la matriz de datos, **A** es el vector de incógnitas, correspondiente a los coeficientes de predicción, y **E** es el vector de términos independientes, correspondiente a los errores de predicción.

El error cuadrático total de predicción E es simplemente

$$E = \sum_n e^2(n) = \sum_n \left(x(n) + \sum_{k=1}^{p} a_k x(n-k) \right)^2, \qquad (3.15)$$

donde el rango de sumatorio en n no se ha especificado por las razones que se verán más adelante.

Para minimizar E, basta con derivar (3.15) con respecto a los coeficientes de predicción e igualar a 0. El resultado es

$$\sum_{k=1}^{p} a_k \sum_{n} x(n-k)x(n-i) = - \sum_{n} x(n)x(n-i) \qquad i = 1,\dots p, \qquad (3.16)$$

con error cuadrático total mínimo E_p

$$E_p = \sum_{n} x^2(n) + \sum_{k=1}^{p} a_k \sum_{n} x(n) \quad x(n-k). \qquad (3.17)$$

Las expresiones (3.16) y (3.17) pueden escribirse matricialmente de la forma

$$(\mathbf{X}_i^T \mathbf{X}_i) \mathbf{A} = \begin{pmatrix} E_p \\ 0 \\ . \\ . \\ 0 \end{pmatrix} \qquad (3.18)$$

donde el subíndice i selecciona una de las matrices $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ y $\mathbf{X}_4$ indicadas en (3.13). Esta selección depende del rango del sumatorio en n de las expresiones (3.15)-(3.17), es decir del rango de valores de n en que se minimiza el error cuadrático total. Si este rango es n =1,... N +p, se selecciona la matriz $\mathbf{X}_1$; si el rango es n=p+1,... N, se selecciona la matriz $\mathbf{X}_2$; etc.

La selección de la matriz $\mathbf{X}_1$ da lugar al método de autocorrelación y la selección de la matriz $\mathbf{X}_2$ da lugar al método de covarianza. Ambos métodos serán revisados en los dos siguientes apartados por ser los más usados. La selección de las matrices $\mathbf{X}_3$ y $\mathbf{X}_4$ dan lugar a los métodos de preenventanado y postenventado, respectivamente.

3.2.1.1. METODO DE AUTOCORRELACION. ECUACIONES DE YULE-WALKER (YWE)

Si denotamos con r(m) el estimador sesgado clásico de la autocorrelación para una secuencia finita de muestras x(n), n = 1,... N, omitiendo el factor constante 1/N, es decir,

$$r(m) = \sum_{n=1}^{N-m} x(n+m)x(n), \qquad (3.19)$$

la expresión (3.18), en el caso de seleccionar la matriz $\mathbf{X}_1$, puede escribirse como

$$\begin{pmatrix} r(0) & r(1) & r(2) & \dots & r(p) \\ r(1) & r(0) & r(1) & \dots & r(p-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} E_p \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad (3.20)$$

donde se ha tenido en cuenta que la secuencia de autocorrelación r(m) es una secuencia par. De ahí el nombre de método de autocorrelación.

El sistema de ecuaciones (3.20) es conocido como ecuaciones de Yule-Walker (YWE, *Yule-Walker Equations*). Debido a que la matriz de autocorrelaciones del sistema es simétrica y Toeplitz, es decir, los elementos de cualquier diagonal son idénticos, este sistema puede resolverse de forma eficiente mediante el algoritmo de Levinson-Durbin (requiere sólo un número de operaciones del orden de p^2, mientras que el método de eliminación de Gauss requiere del orden de p^3 operaciones).

El algoritmo de Levinson-Durbin calcula de forma recursiva los predictores para orden desde j = 1 hasta p. Es decir, calcula los conjuntos $\{a_{11}, E_1\}$, $\{a_{21}, a_{22}, E_2\}$,... $\{a_{p1}, a_{p2}, \dots a_{pp}, E_p\}$, donde el primer subíndice de los coeficientes de predicción indica el orden.

La inicialización del algoritmo es

$$a_{11} = -r(1)/r(0) \qquad (3.21)$$

$$E_1 = (1 - a_{11}^2)\, r(0) \qquad (3.22)$$

y la recursión para j = 1,... p viene dada por

$$a_{jj} = - \left[r(j) + \sum_{l=1}^{j-1} a_{j-1,l}\ r(j-l) \right] / E^2_{j-1} \qquad (3.23)$$

$$a_{ji} = a_{j-1,i} + a_{jj}\ a_{j-1,j-i} \qquad (3.24)$$

$$E^2_j = (1 - a^2_{jj})\ E^2_{j-1}. \qquad (3.25)$$

La solución final es, pues,

$$a_j = a_{pj} \qquad (3.26)$$

Salvo en el caso de que la señal x(n) responda exactamente a un modelo todo polos, cosa que no ocurre cuando se trabaja con señal de voz, se demuestra fácilmente a partir de las expresiones anteriores que el error cuadrático total mínimo disminuye al aumentar el orden. Por tanto, los términos E_j de este algoritmo pueden ser de gran ayuda para seleccionar el orden de predicción.

El método de autocorrelación supone una señal estacionaria e infinita que ha sido enventanada. Notar en la representación esquemática del método de la figura 3.7 que se minimiza el error de predicción e(n) desde n =1 a N+p y para ello se suponen nulas las muestras de x(n) para n < 1 y n > N. La forma de esta ventana afecta a los valores de los coeficientes de predicción y, por tanto, a la consiguiente estimación espectral. En particular, si se usa la ventana rectangular, implícita en la formulación anterior, los lóbulos laterales de su transformada enmascaran frecuentemente los formantes más altos. Por ello, es necesaria la aplicación de una ventana sobre la señal que suprima en lo posible los lóbulos laterales. La ventana de Hamming es es las más usada y es la que se ha elegido para las pruebas experimentales de este trabajo.

Otra consecuencia importante del enventanado de la señal es que pueden aparecer problemas de resolución si la longitud de la trama de señal no es lo suficientemente grande. Experimentalmente, se ha comprobado que en el caso de sonidos sonoros la trama de señal ha de abarcar varios períodos para obtener resultados fiables. A una frecuencia de muestreo de 8 kHz, que es la utilizada en las pruebas experimentales de este trabajo, se suelen utilizar valores de N comprendidos entre 100 y 400. En este trabajo se han utilizado tramas de voz de 240 muestras, que se corresponden con una duración temporal de 30 ms.

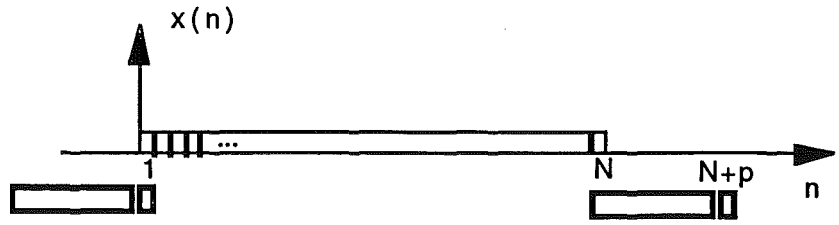


Fig. 3.7. Representación esquemática del método de autocorrelación

3.2.1.2. METODO DE COVARIANZA

Si se selecciona la matriz $\mathbf{X}_2$ en la expresión (3.18), se obtiene un sistema de ecuaciones en que la matriz de coeficientes tiene las propiedades de una matriz de covarianza.

La simetría de esta matriz permite la utilización de algoritmos eficientes. Entre ellos, el más utilizado es la descomposición de Cholesky. Sin embargo, estos algoritmos no son tan eficientes como el de Levinson-Durbin, pues este aprovecha, además de la propiedad de simetría, el carácter Toeplitz de la matriz del método de autocorrelación. Así, por ejemplo, para p = 10, el algoritmo de Levinson-Durbin es tres veces más eficiente computacionalmente que el de Cholesky.

En el método de covarianza no existe el problema de enventanado que se producía en el método de autocorrelación. Notar en la representación esquemática del método de la figura 3.8 que se minimiza el error de predicción e(n) desde n = p+1 a N y para ello no se ha de suponer nula ninguna muestra de x(n). Por ello, pueden obtenerse estimaciones más precisas con tramas más cortas. Sin embargo, en la mayoría de las aplicaciones se toman tramas de longitud comparable a las que se toman en el método de autocorrelación. Ello es debido a que al tomar tramas cortas en el caso de sonidos sonoros la posición relativa del máximo del período con respecto al inicio de la trama puede provocar problemas importantes.

En procesado del habla, la experiencia demuestra que el método de autocorrelación proporciona mejores resultados con sonidos fricativos y el de covarianza para sonidos periódicos. Cuando N aumenta los dos métodos tienden a aproximarse.

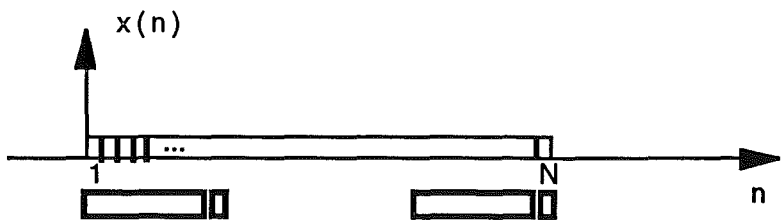


Fig. 3.8. Representación esquemática del método de covarianza

En el sistema de reconocimiento básico de las pruebas experimentales realizadas en este trabajo, se ha utilizado el método de autocorrelación, en lugar del de covarianza, debido a su mayor eficiencia computacional.

3.2.2. MODELADO ESPECTRAL

Del modelo de generación de señal correspondiente a la técnica de predicción lineal representado en la figura 3.6, se obtiene que el espectro de la señal $S_{xx}(\omega)$ viene dado por la expresión [Mak75]

$$S_{xx}(\omega) = \frac{S_{ee}(\omega)}{|A(e^{j\omega})|^2}, \qquad (3.27)$$

donde $S_{ee}(\omega)$ es el espectro del error de predicción e(n) y $A(e^{j\omega})$ es la respuesta frecuencial del filtro de error de predicción.

El modelado espectral asociado a la técnica de predicción lineal consiste en aproximar este espectro por el módulo al cuadrado de la respuesta frecuencial del filtro todo-polos H(z), es decir,

$$\hat{S}_{xx}(\omega) = \frac{G^2}{|A(e^{j\omega})|^2}, \qquad (3.28)$$

donde $\hat{S}_{xx}(\omega)$ es la aproximación de $S_{xx}(\omega)$ dada por la predicción lineal.

Comparando (3.27) y (3.28), se observa que el espectro del error $S_{ee}(\omega)$ se modela por un espectro plano igual a G^2. Es decir, la señal de error e(n) se aproxima por otra señal cuyo espectro es plano, como por ejemplo ruido blanco o un impulso. En el caso de predicción lineal de la señal voz, el ruido blanco corresponde a la excitación de los sonidos sordos y el impulso corresponde a la de los sonidos sonoros. Nótese que en en el caso de los sonidos sonoros se ha perdido la periodicidad de la señal, ya que se ha eliminado la estructura final del espectro.

Por otro lado, teniendo en cuenta que el error cuadrático total puede E escribirse como

$$E = \frac{1}{2\pi} \int_{-p}^{p} S_{ee}(\omega)\, d\omega \qquad (3.29)$$

y combinando las expresiones (3.27) y (3.28), podemos expresar E como

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{S_{xx}(\omega)}{\hat{S}_{xx}(\omega)}\, d\omega. \qquad (3.30)$$

Por tanto, minimizar el error cuadrático total E es equivalente a minimizar la integral del cociente entre el espectro de la señal y su aproximación. Debido a ello, los casos en que $S_{xx}(\omega) > \hat{S}_{xx}(\omega)$ contribuirán más al error que los casos en que $S_{xx}(\omega) < \hat{S}_{xx}(\omega)$. Esto conduce a que $\hat{S}_{xx}(\omega)$ tienda a seguir los picos de $S_{xx}(\omega)$ más que los valles. En particular, si $S_{xx}(\omega)$ es el espectro de una señal de voz, $\hat{S}_{xx}(\omega)$ intenta aproximar la envolvente espectral (ver figura 3.3), es decir, tenderá a aproximar $S_{xx}(\omega)$ de forma más exacta alrededor de los picos de los formantes.

Por tanto, mediante la técnica de predicción lineal se consigue separar eficientemente la estructura fina del espectro de la señal de voz de su envolvente, correspondientes a la excitación u(n) y al filtro H(z), respectivamente, del modelo simplificado de producción de voz de la figura 3.5. La estructura fina del espectro es asociada a $S_{ee}(\omega)$ (recordar que e(n)=Gu(n)) y la envolvente es asociada a $\hat{S}_{xx}(\omega)$

Para ilustrar el modelado espectral de predicción lineal, la figura 3.9 representa, para p =14: a) trama de señal de voz enventanada correspondiente a la

vocal /a/; b) error de predicción; c) espectro de la señal y espectro del modelo; d) espectro de la señal de error. Se ha utilizado el método de autocorrelación con ventana de Hamming sobre una trama de 200 puntos, resultado de muestrear la señal de voz a una frecuencia de 10 kHz. En la figura puede observarse la aproximación de la envolvente del espectro de la señal realizada por el modelo y la planicidad del espectro de error de predicción.

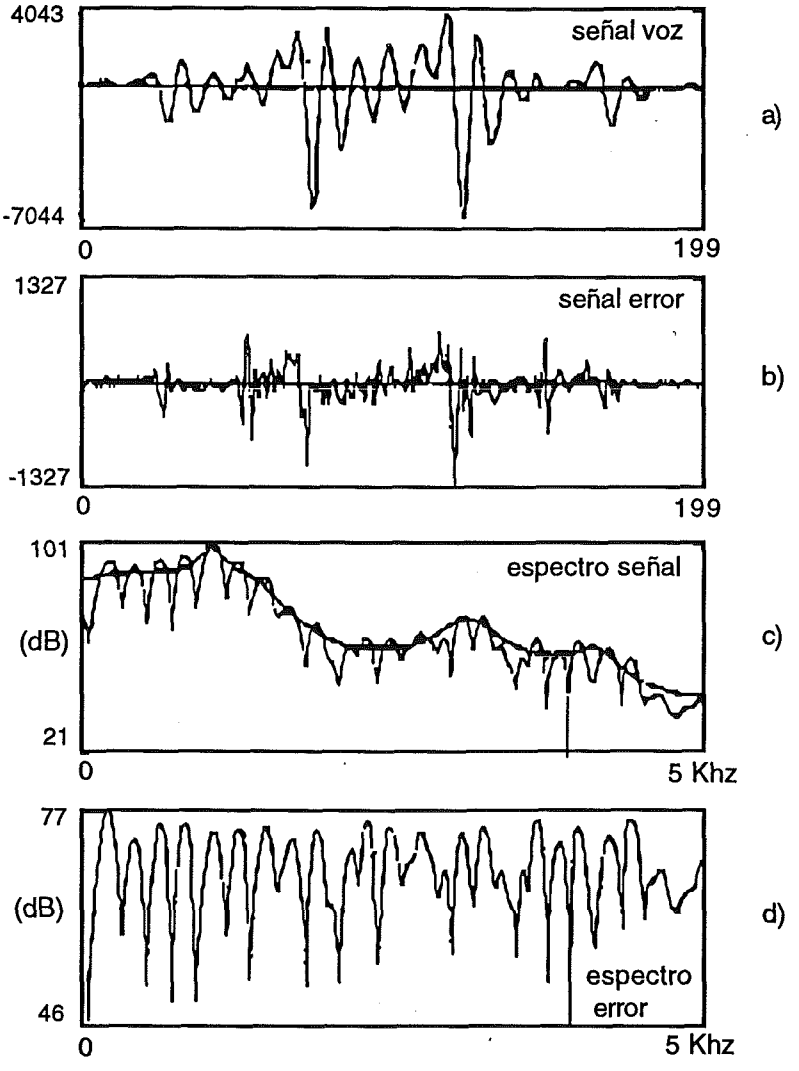


Fig. 3.9. Señales y espectros obtenidos del análisis LPC para la vocal /a/.

Una cuestión importante en la predicción lineal de la señal de voz es la elección del orden de predicción p necesario para capturar la estructura de formantes de la señal. Si se escoge un orden demasiado bajo, se obtiene un espectro muy suavizado en el que se puede haber perdido información de algunos formantes. Es razonable escoger un orden p igual a la frecuencia de muestreo expresada en kHz, debido a que el tiempo invertido por el sonido en recorrer dos veces la longitud de un tracto vocal medio es aproximadamente 1 ms, y en algunos casos se añaden algunos términos más para modelar otros efectos.

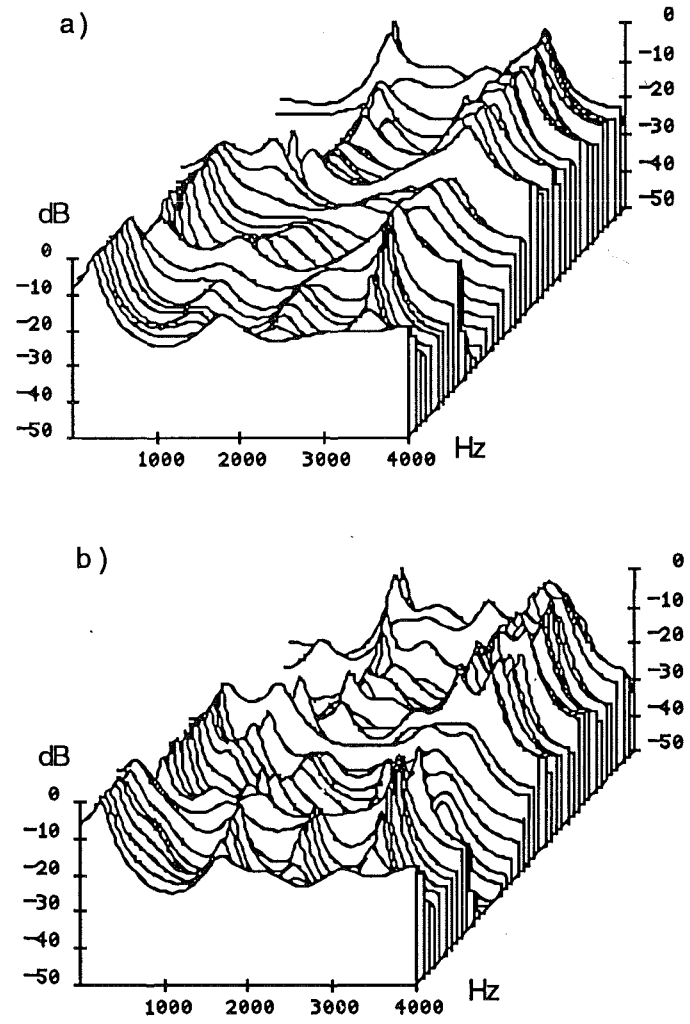


Fig. 3.10. Influencia del orden de predicción: a) p=8, b) p= 12

En la figura 3.10 se representa la secuencia de espectros obtenida aplicando el método de autocorrelación sobre tramas de 240 muestras, tomadas con un desplazamiento de 120 muestras, para órdenes del modelo predicción 8 y 12. La señal de voz, muestreada a 8 kHz, corresponde al dígito 0 pronunciado en catalán. Como puede observarse, al aumentar el orden de predicción aparecen nuevos picos en la envolvente espectral correspondientes a los nuevos polos de filtro H(z). En el sistema básico de reconocimiento utilizado en las pruebas experimentales realizadas en este trabajo la frecuencia de muestreo es de 8 kHz, para abarcar el canal telefónico, y el orden de predicción también es 8.

Por último, cabe destacar que antes de realizar el análisis de predicción lineal sobre la señal de voz esta suele ser filtrada paso alto mediante un filtro de función de transferencia

$$H_p(z) = 1 - a\, z^{-1} \qquad (3.31)$$

Normalmente se usan valores de a entre 0.9 y 1 (en el sistema básico de reconocimiento utilizado en las pruebas experimentales de este trabajo μ es igual a 0.95). La razón principal es filtrado, que recibe el nombre de preénfasis, es reducir el rango dinámico del espectro de la señal de voz, lo cual disminuye los problemas numéricos en la implementación práctica.

3.3. PREDICCION LINEAL EN PRESENCIA DE RUIDO

Un importante problema de las técnicas de predicción lineal vistas en el apartado anterior es su sensibilidad al ruido aditivo, es decir, se produce una importante degradación de la calidad de las estimaciones espectrales obtenidas cuando la señal está contaminada de ruido. Este hecho limita su utilización en reconocimiento del habla en entornos ruidosos.

Si el ruido es de carácter periódico, por ejemplo, el procedente de motores, el predictor intentará modelar los picos espectrales correspondientes a las periodicidades del ruido ya que, como se ha visto en el apartado anterior, la predicción lineal tiende a favorecer los picos del espectro sobre los valles.

En el caso de ruido blanco, este reduce el rango dinámico del espectro, es decir, tiende a aplanarlo. Debido a ello, los polos del modelo de predicción lineal tienden a trasladarse hacia el origen del plano z [Kay79]. Además del suavizado excesivo del espectro del modelo, se observa también un desplazamiento de los picos, que en el caso de la señal de voz se corresponden con los formantes. Relaciones de señal-ruido bajas, por ejemplo, por debajo de 5 o 10 dB, pueden causar serias distorsiones en el modelado espectral.

Estos efectos pueden observarse claramente en la figura 3.11. En ella están representados los espectros LPC de orden 12 de un segmento sonoro de señal de voz, parte estacionaria de la vocal /o/, en condiciones supuestamente libres de ruido (línea continua) y en presencia de ruido blanco aditivo de igual potencia que la señal (línea de puntos). Estos espectros han sido obtenidos aplicando el método de autocorrelación a una trama de 240 muestras (frecuencia de muestreo 8 kHz).

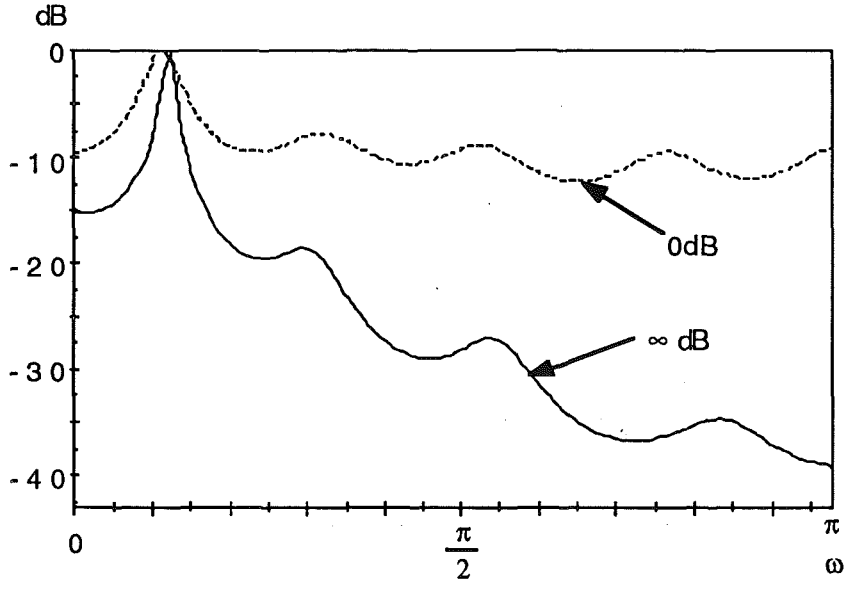


Fig. 3.11. Efecto del ruido blanco aditivo sobre el espectro LPC

La razón más importante de esta degradación es que el modelo todo-polos supuesto por estas técnicas deja de ser válido en presencia de ruido. Así, por ejemplo, si x(n) es un proceso cuyo espectro viene dado por la expresión (3.28) y w(n) es

ruido blanco aditivo e incorrelado con dicho proceso de potencia σ^2, el espectro del proceso contaminado

$$y(n) = x(n) + w(n) \tag{3.32}$$

tiene la expresión

$$S_{yy}(\omega) = \frac{G^2}{|A(e^{j\omega})|^2} + \sigma^2 = \frac{G^2+\sigma^2|A(e^{j\omega})|^2}{|A(e^{j\omega})|^2}. \tag{3.33}$$

Por tanto, y(n) es un proceso cuyo espectro tiene ceros y polos, en particular, igual número de ceros y polos. Como se verá en el apartado 3.3.1, a los procesos como x(n) cuyo espectro es todo-polos se le denomina autorregresivos (abreviadamente, AR) y a los procesos como y(n) cuyo espectro es racional se les denomina procesos ARMA.

No hay soluciones simples a este problema. Aparte de la posibilidad de realizar un procesado de la señal para atenuar el ruido, que ya ha sido comentada en el apartado 2.3.2 de esta memoria, se han propuesto tres aproximaciones básicas:

a) compensación o bien de las estimaciones de autocorrelación o de los coeficientes de reflexión (coeficientes a_{jj}, j = 1,... p, en el algoritmo de Levinson-Durbin),

b) utilización de un orden de predicción alto,

c) uso de métodos de estimación espectral para procesos ARMA.

En cuanto a la primera aproximación, la relación entre las secuencias de autocorrelación de la señal contaminada $r_{yy}(n)$ y de la señal limpia $r_{xx}(n)$ en el caso de ruido blanco aditivo e incorrelado de potencia σ^2 es

$$r_{yy}(n) = r_{xx}(n) + \sigma^2\,\delta(n). \tag{3.34}$$

Por tanto, la corrección de la estimación del valor en el origen de la autocorrelación de la señal de voz sustrayendo una estimación de la potencia de ruido puede servir para eliminar los efectos del ruido. Esta aproximación es atractiva por su simplicidad en el caso de realizare predicción lineal mediante el método de autocorrelación. Sin embargo, una seria deficiencia de esta técnica es que la estimación de σ^2 no es segura. Si se

extrae demasiada potencia de ruido, el espectro estimado exhibirá picos más abruptos que el espectro real. Además, la sustracción de una cantidad errónea de potencia de ruido puede dar a lugar a una secuencia que no sea de autocorrelación y, por tanto, a filtros H(z) no estables. Para evitar este problema, Kay [Kay80] propuso un método de compensación de los coeficientes de reflexión, que ofrece la ventaja de garantizar la estabilidad. No obstante, aunque este tipo de técnicas pueden reducir el sesgo de la estimación, tienden a incrementar la varianza de la misma.

La utilización de un orden de predicción más alto que el orden del modelo correspondiente a la señal limpia se basa en el hecho de cualquier proceso puede modelarse exactamente utilizando un orden de predicción suficientemente alto. Si el proceso es AR, este orden de predicción es finito y se corresponde con el número de polos del modelo; en caso contrario, se necesitaría un orden de predicción infinito y a medida que aumenta el orden usado se aproxima mejor el espectro del proceso. Por tanto, en el caso de señal ruidosa, teóricamente debería utilizarse el mayor orden de predicción posible. No obstante, en la práctica, si se utiliza un orden de predicción demasiado alto aparecen picos espurios debidos a los polos extra generados por los errores de estimación. En las pruebas experimentales presentadas en el capítulo 6, se ha variado el orden de predicción lineal y se ha estudiado su influencia en la tasa de reconocimiento en el caso de señal de voz limpia y ruidosa. Se ha observado que el reconocimiento de habla ruidosa requiere órdenes de predicción superiores a los utilizados en reconocimiento de habla libre de ruido.

Con respecto a las técnicas de estimación espectral para procesos ARMA, una solución adecuada al problema del ruido sería la estimación de máxima verosimilitud para este tipo de procesos. Sin embargo, este procedimiento conduce a un conjunto de ecuaciones altamente no-lineales. Una solución subóptima a las ecuaciones de máxima verosimilitud para el caso de un proceso AR en ruido blanco conduce a un filtrado iterativo [Lim78]. Ya fuera del ámbito de la estimación de máxima verosimilitud, las formulaciones que intentan hallar simultáneamente todos los parámetros de un proceso ARMA dan lugar también a métodos iterativos [Kay87]. Debido principalmente al elevado coste computacional de estas técnicas, se recurre a la estimación separada de los parámetros AR del proceso ARMA. La aproximación más básica es el uso de las ecuaciones de Yule-Walker de orden superior (HOYWE, *High-Order Yule-Walker Equations*) [Ger70] [Don78]. Aunque simple de implementar, este método sólo obtiene buenos resultados en tramas largas y/o relaciones señal-ruido altas. Estos problemas intentan subsanarse usando un sistema sobredeterminado de ecuaciones de Yule-Walker

de orden superior extendidas [Cad82], que se denotarán en esta memoria con las siglas OHOYWE (*Overdetermined High Order Yule-Walker Equations*).

A continuación, se revisara en el apartado 3.3.1 la teoría correspondiente al modelado de procesos mediante funciones de transferencia racionales, que da lugar a los modelos AR, MA y ARMA. Seguidamente, en los apartados 3.3.2 y 3.3.3 se abordarán los métodos de estimación basados en el uso de las HOYWE y OHOYWE, respectivamente. Finalmente, se describirá en el apartado 3.3.4 el uso de un sistema sobredeterminado de ecuaciones de Yule-Walker extendidas para las estimación fiable de los parámetros de un proceso AR que se denotarán con las siglas OYWE (*Overdetermined Yule-Walker Equations*).

3.3.1. MODELADO AR, MA Y ARMA

Muchos procesos discretos encontrados en la práctica pueden aproximarse mediante un modelo de función de transferencia racional. En este modelo, la señal observada se modela como la salida x(n) de un filtro causal de función de transferencia racional H(z)

$$H(z) = G\,\frac{B(z)}{A(z)} = G\,\frac{1+\sum_{k=1}^{q} b_k z^{-k}}{1+\sum_{k=1}^{p} a_k z^{-k}} \qquad (3.35)$$

excitado por una entrada u(n). De este modo, la relación entre x(n) y u(n) viene dada por la ecuación en diferencias

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + G\sum_{k=0}^{q} b_k u(n-k), \qquad (3.36)$$

donde se ha se considerado $b_0=1$.

La secuencia u(n) normalmente es desconocida. Suele suponerse que se trata de ruido blanco de media nula y varianza unidad (la ganancia del filtro permite escalar la energía de la señal). Por tanto su espectro tiene valor un valor constate unidad.

Considerando este tipo de excitación u(n), la ecuación (3.36) determina un proceso ARMA (*AutoRegressive-Moving Average*), cuyo espectro será

$$S_{xx}(\omega) = G^2 \frac{|B(e^{j\omega})|^2}{|A(e^{j\omega})|^2}. \qquad (3.37)$$

Usualmente se utiliza la notación ARMA(p,q) para indicar un proceso ARMA en que el polinomio A(z) es de orden p y el polinomio B(z) es de orden q.

En el caso particular de que todos los coeficientes a_k, k = 1,... p, sean nulos, entonces

$$x(n) = G \sum_{k=0}^{q} b_k u(n-k) \qquad (3.38)$$

$$S_{xx}(\omega) = G^2 \, |B(e^{j\omega})|^2. \qquad (3.39)$$

Se dice, entonces, que el proceso es MA (*Moving Average*) de orden q y se denota como MA(q).

Cuando todos los coeficientes b_k son nulos, excepto $b_0 = 1$, se cumple

$$x(n) = - \sum_{k=1}^{p} a_k x(n-k) + Gu(n), \qquad (3.40)$$

$$S_{xx}(\omega) = \frac{G^2}{|A(e^{j\omega})|^2}. \qquad (3.41)$$

Entonces, se dice que el proceso es AR (*AutoRegressive*) de orden p y se denota como AR(p).

El modelo espectral todo-polos asociado a la predicción lineal clásica de orden p (3.28) es equivalente al espectro de un proceso AR del mismo orden p (3.41). Ello es debido a que el modelo espectral de predicción lineal aproxima el error de predicción e(n) = Gu(n) por una señal de espectro plano y, en este caso, los modelos de generación de señal para ambos casos, (3.9) y (3.40), coinciden.

Por tanto, se puede enfocar el modelado espectral de la señal de voz realizado por la predicción lineal desde el punto de vista del modelado AR de un proceso. En la

práctica, todas las técnicas de predicción lineal son aplicables a la estimación de los parámetros autorregresivos a_k en el modelado AR, equivalentes a los coeficientes de predicción en predicción lineal. Seguidamente se verá cómo el método de autocorrelación de predicción lineal puede también derivarse utilizando conceptos de modelado AR exclusivamente.

Para un proceso real AR es fácil establecer la siguiente relación entre las autocorrelaciones exactas r(m) del proceso y los parámetros autorregresivos a_k

$$r(m) = - \sum_{k=1}^{p} a_k \; r(m-k) \qquad m > 0 \qquad (3.42)$$

$$r(0) = - \sum_{k=1}^{p} a_k \; r(-k) + G^2 \qquad (3.43)$$

$$r(m) = r(-m) \qquad m < 0 \qquad (3.44)$$

Escribiendo matricialmente esta relación para $0 \leq m \leq p$, se obtiene

$$\begin{pmatrix} r(0) & r(1) & r(2) & \dots & r(p) \\ r(1) & r(0) & r(1) & \dots & r(p-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} G^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (3.45)$$

Esta ecuaciones permiten obtener los parámetros autorregresivos a_k del modelo AR a partir de los valores exactos de la secuencia de autocorrelación desde m = 0 a p. Como puede observarse, estas son las ecuaciones de Yule-Walker de la expresión (3.20), cambiando E_p por G^2, y pueden resolverse eficientemente mediante el algoritmo de Levinson-Durbin (3.21)-(3.26) debido a que la matriz de autocorrelaciones es simétrica y Toeplitz.

En general, los valores exactos de la autocorrelación no son conocidos y es necesario estimar los valores de la autocorrelación a partir de la señal x(n) para construir la matriz de autocorrelaciones del sistema (3.45). Según la forma de estimar las autocorrelaciones, se obtendrán diferentes estimaciones de los parámetros autorregresivos. En particular, si se elige el estimador sesgado clásico de autocorrelación (3.19) este método equivale al método de autocorrelación de predicción lineal descrito en el apartado 3.2.1.1.

3.3.2. ECUACIONES DE YULE-WALKER DE ORDEN SUPERIOR (HOYWE)

Se ha visto en (3.33) que un proceso AR(p) en presencia de ruido blanco aditivo e incorrelado es equivalente a un proceso ARMA(p,p) con los mismos parámetros autorregresivos a_k. Por tanto, la estimación de los parámetros de un proceso AR(p) en presencia de ruido blanco aditivo e incorrelado se reduce a la estimación de los parámetros autorregresivos de un proceso ARMA(p,p)

Como ya se ha discutido, una estimación conjunta de todos los parámetros de un proceso ARMA conduce a algoritmos iterativos de elevado coste computacional, por lo cual se suele realizar una estimación subóptima separada de los parámetros autorregresivos a_k y los *moving average* b_k. Seguidamente, se describirá la aproximación básica para estimar los parámetros autorregresivos a_k, que consiste en la resolución las llamadas ecuaciones de Yule-Walker de orden superior (HOYWE, *High Order Yule-Walker Equations*) [Ger70] [Don78].

Para un proceso real ARMA(p,q), cuyo espectro responde a la expresión (3.37), es fácil establecer la siguiente relación entre las autocorrelaciones exactas r(m) del proceso y los parámetros a_k y b_k

$$r(m) = -\sum_{k=1}^{p} a_k r(m-k) \qquad m>q \qquad (3.46)$$

$$r(m) = -\sum_{k=1}^{p} a_k r(m-k) + \sum_{k=1}^{q} b_k h(k-m) \qquad 0 \le m \le q \qquad (3.47)$$

$$r(m) = r(-m) \qquad m < 0, \qquad (3.48)$$

donde h(n) es la respuesta impulsional del filtro H(z).

Escribiendo matricialmente la relación (3.46) para m = q+1,... q+p, se obtiene

$$\begin{pmatrix} r(q+1) & r(q) & r(p-1) & \dots & r(q-p+1) \\ r(q+2) & r(q+1) & r(q) & \dots & r(q-p+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(q+p) & r(q+p-1) & r(q+p-2) & \dots & r(q) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (3.49)$$

Esta ecuaciones son conocidas como ecuaciones de Yule-Walker de orden superior (HOYWE) y permiten obtener los parámetros autorregresivos a_k del modelo ARMA(p,q) a partir de los valores exactos de la secuencia de autocorrelación desde m = q-p+1 a q+p. En la práctica, cuando no se dispone de los valores exactos de la autocorrelación se estiman a partir de la señal.

En el caso de un proceso ARMA(p,p) estas ecuaciones toman la forma

$$\begin{pmatrix} r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ r(p+2) & r(p+1) & r(p) & \dots & r(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(2p) & r(2p-1) & r(2p-2) & \dots & r(p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (3.50)$$

Por tanto, utilizando un estimador adecuado de la autocorrelación a partir de la señal, estas ecuaciones pueden usarse para estimar los parámetros autorregresivos a_k de un un proceso AR(p) en presencia de ruido blanco aditivo.

Estas mismas ecuaciones (3.50) podrían haberse derivado directamente a partir de la evaluación de m = p+1 a 2p de la expresión (3.42), que relaciona las autocorrelaciones teóricas de un proceso AR(p) con los parámetros a_k para m > 0. Como en un proceso AR(p) en presencia de ruido blanco aditivo el único valor de la autocorrelación contaminado es r(0) y en el sistema de ecuaciones (3.50) este valor no aparece, este sistema se ha de cumplir para un proceso AR(p) tanto en ausencia como en presencia de este tipo de ruido.

Estos métodos son computacionalmente muy atractivos. Sin embargo, las estimaciones de los parámetros autorregresivos de un proceso ARMA(p,q) utilizando el sistema de ecuaciones (3.49) son, en general, de baja calidad debido en gran parte a la varianza de las estimaciones de los valores de la autocorrelación, que aumenta con el índice m por disminuir el número de datos que intervienen en dicha estimación. En el caso de aplicación de las ecuaciones (3.50) a un proceso AR(p) en presencia de ruido, sólo se obtienen resultados razonables en tramas largas y/o relaciones señal-ruido altas.

3.3.3. ECUACIONES SOBREDETERMINADAS DE YULE-WALKER DE ORDEN SUPERIOR (OHOYWE)

Una aproximación alternativa [Cad82] para mejorar la estimación de los p parámetros autorregresivos de un proceso ARMA(p,q) es el uso de un sistema sobredeterminado de más de p ecuaciones obtenidas evaluando (3.46) para valores de m mayores que q consecutivos, m = q+1, q+2,...

Esta aproximación está basada en el hecho de que en las HOYWE sólo intervienen los valores estimados de la autocorrelación de m = p-q+1 a q+p y, por tanto, los parámetros autorregresivos obtenidos dependen totalmente de las estimaciones de la secuencia de autocorrelación para estos valores de m, que presentan un cierto error de estimación. Estos errores de estimación pueden compensarse mediante la utilización de más del número mínimo p de ecuaciones, con lo cual se hace intervenir en la obtención de los parámetros autorregresivos un conjunto mayor de valores estimados de autocorrelación.

Suponiendo que M, tal que M-q > p, es el mayor índice para el que la autocorrelación puede estimarse con cierta fiabilidad, puede construirse el siguiente sistema sobredeterminado de M-q ecuaciones y p incógnitas evaluando (3.46) desde m = q+1 hasta M

$$\begin{pmatrix} r(q+1) & r(q) & r(p-1) & \dots & r(q-p+1) \\ r(q+2) & r(q+1) & r(q) & \dots & r(q-p+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p+p) & r(p+p-1) & r(q+p-2) & \dots & r(p) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(q+1) \\ \varepsilon(q+2) \\ \vdots \\ \varepsilon(q+p) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix}, \quad (3.51)$$

donde $\varepsilon(m)$ es el error asociado a la estimación de las autocorrelaciones. Estas ecuaciones se denotarán en esta memoria con las siglas OHOYWE (*Overdetermined High Order Yule-Walker Equations*) y pueden servir para la obtención de los parámetros autorregresivos de un proceso ARMA(p,q) utilizando un estimador adecuado de los valores de la autocorrelación.

La aproximación básica para la resolución del sistema (3.51) es la de mínimos cuadrados, que consiste en minimizar el error cuadrático

$$E = \sum_{m=q+1}^{M} |\varepsilon(m)|^2 \qquad (3.52)$$

con respecto a los p parámetros autorregresivos a_k. Esta aproximación es equivalente a la aplicación del método de covarianza de predicción lineal sobre la secuencia de autocorrelación r(q-p+1),... r(M), en lugar de la señal x(1),... x(N).

Para disminuir el efecto del incremento de la varianza asociada a la estimación de los valores de la autocorrelación al aumentar el índice m, se ha propuesto [Fri85] la minimización de un error cuadrático ponderado

$$E = \sum_{m=q+1}^{M} w(m)|\varepsilon(m)|^2, \qquad (3.53)$$

donde w(m) es una secuencia de ponderación decreciente con m. Sin embargo, la elección de la ponderación w(n) adecuada en cada caso es difícil.

También se ha propuesto [Cad82] la utilización de técnicas de descomposición en valores singulares (SVD, *Singular Value Decomposition*) para la resolución del sistema (3.51).

En el caso de un proceso ARMA(p,p) el sistema (3.51) adopta la forma

$$\begin{pmatrix} r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ r(p+2) & r(p+1) & r(p) & \dots & r(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(2p) & r(2p-1) & r(2p-2) & \dots & r(p) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(p+1) \\ \varepsilon(p+2) \\ \vdots \\ \varepsilon(2p) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix} \qquad (3.54)$$

y puede servir para la estimación de los parámetros regresivos de un proceso AR(p) en presencia de ruido blanco aditivo utilizando un estimador adecuado de los valores de la autocorrelación.

Como puede observarse, este sistema de ecuaciones consiste en una extensión del sistema (3.50). Una posible justificación de la mejora que puede suponer en la práctica la utilización del sistema (3.54), en lugar del sistema (3.50), para la

estimación de un proceso AR(p) en presencia de ruido blanco aditivo es que los valores de autocorrelación alejados del origen son más robustos al ruido blanco que los cercanos al origen (en la práctica, el ruido no es idealmente blanco y la estimación se realiza en un intervalo finito, por lo que este no sólo afecta a r(0) sino que contamina los valores de la autocorrelación r(m) de forma decreciente con el índice m).

3.3.4. ECUACIONES SOBREDETERMINADAS DE YULE-WALKER (OYWE)

Si no interesa el cálculo de la ganancia G del modelo, como es el caso de la aplicación a reconocimiento del habla, se puede suprimir la primera de las ecuaciones del sistema de ecuaciones de Yule-Walker (3.45). Por tanto, se puede escribir

$$\begin{pmatrix} r(1) & r(0) & r(1) & \dots & r(p-1) \\ r(2) & r(1) & r(0) & \dots & r(p-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (3.55)$$

Teniendo en cuenta las justificaciones dadas en el apartado anterior para el uso del sistema (3.51) para la estimación fiable de los parámetros autorregresivos de un proceso ARMA(p,q) y del sistema (3.54) en el caso de un proceso AR(p) en presencia de ruido blanco aditivo, se puede extender el sistema de ecuaciones (3.55) al siguiente sistema

$$\begin{pmatrix} r(1) & r(0) & r(1) & \dots & r(p-1) \\ r(2) & r(1) & r(0) & \dots & r(p-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(p) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix} , \qquad (3.56)$$

que puede utilizarse como un método fiable de estimación de los parámetros autorregresivos de un proceso AR(p) utilizando un estimador adecuado de los valores de la autocorrelación. Estas ecuaciones se denotarán en esta memoria con el nombre de ecuaciones sobredeterminadas de Yule-Walker (OYWE, *Overdetermined Yule-Walker Equations*).

3.4. INTERPRETACION COMO PREDICCION LINEAL DE LA SECUENCIA DE AUTOCORRELACION

Como se ha visto en el apartado anterior, si no es necesaria la estimación de la ganancia G del modelo, las ecuaciones de Yule-Walker (YWE) de la expresión (3.45) queda reducido al sistema de ecuaciones (3.55). Como puede observarse en la figura 3.12, este sistema de ecuaciones puede interpretarse como la predicción lineal exacta de orden p de los valores de la secuencia de autocorrelación r(m) desde m = 1 a p utilizando el método de covarianza, ya que no supone ningún enventanado de dicha secuencia.

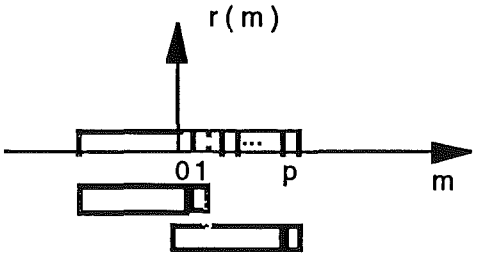


Fig.3.12. Interpretación de las ecuaciones YWE como predicción lineal exacta de la secuencia de autocorrelación de m=1 a p usando el método de covarianza

Análogamente, puede interpretarse el sistema de ecuaciones de Yule-Walker de orden superior HOYWE para un proceso AR(p) en presencia de ruido blanco (3.50) como la predicción lineal exacta de orden p de los valores de la autocorrelación r(m) desde m = p+1 a 2p utilizando el método de covarianza (ver figura 3.13).

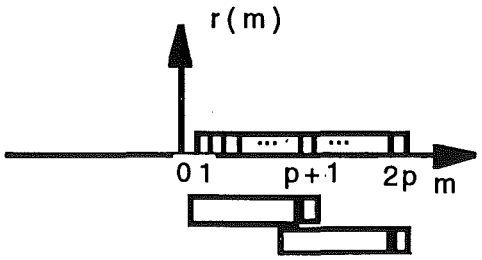


Fig. 3.13. Interpretación de las ecuaciones HOYWE como predicción lineal exacta de la secuencia de autocorrelación de m=p+1 a 2p usando el método de covarianza

También puede interpretarse el sistema de ecuaciones sobredeterminadas de Yule-Walker OYWE para un proceso AR(p) en presencia de ruido blanco (3.56), en el caso de minimizarse el error cuadrático (3.52), como la predicción lineal de los valores de la autocorrelación r(m) desde m = 1 a M utilizando el método de covarianza (ver figura 3.14).

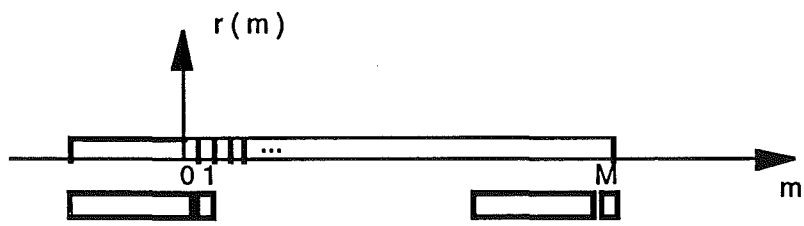


Fig. 3.14. Interpretación de las ecuaciones OYWE como predicción lineal de la secuencia de autocorrelación de m=1 a M usando el método de covarianza

Por último, puede interpretarse el sistema de ecuaciones sobredeterminadas de Yule-Walker de orden superior OHOYWE para un proceso AR(p) en presencia de ruido blanco (3.54), en el caso de minimizarse el error cuadrático (3.52), como la predicción lineal de los valores de la autocorrelación r(m) desde m = p+1 a M utilizando el método de covarianza (ver figura 3.15).

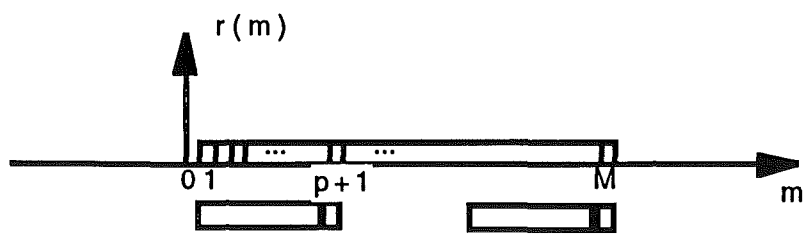


Fig. 3.15. Interpretación de las ecuaciones OHOYWE como predicción lineal de la secuencia de autocorrelación de m=p+1 a M usando el método de covarianza

Por tanto, se ha reducido el problema de la estimación robusta de los parámetros autorregresivos de un proceso AR(p) a la predicción lineal de la secuencia de autocorrelación r(m) de ese proceso mediante la aplicación del método de covarianza a un determinado intervalo de dicha secuencia: de m=1 a p, YWE o LPC clásica; de m=p+1 a 2p, HOYWE; de m = 1 a M, OYWE; y de m=p+1 a M, OHOYWE.

La calidad de las estimaciones espectrales proporcionadas por cada una de estas técnicas dependerá del compromiso robustez al ruido-varianza de las estimaciones de los valores de la secuencia de autocorrelación. Como ya se ha comentado, la varianza de la estimación de los valores r(m) de la secuencia de autocorrelación aumenta con el índice m, ya que disminuye el número de datos que intervienen en dicha estimación. Por otro lado, si el espectro de ruido es plano la robustez de los valores de la autocorrelación aumenta con el índice; en el caso ideal del ruido blanco, este sólo afecta al valor de la autocorrelación en el origen.

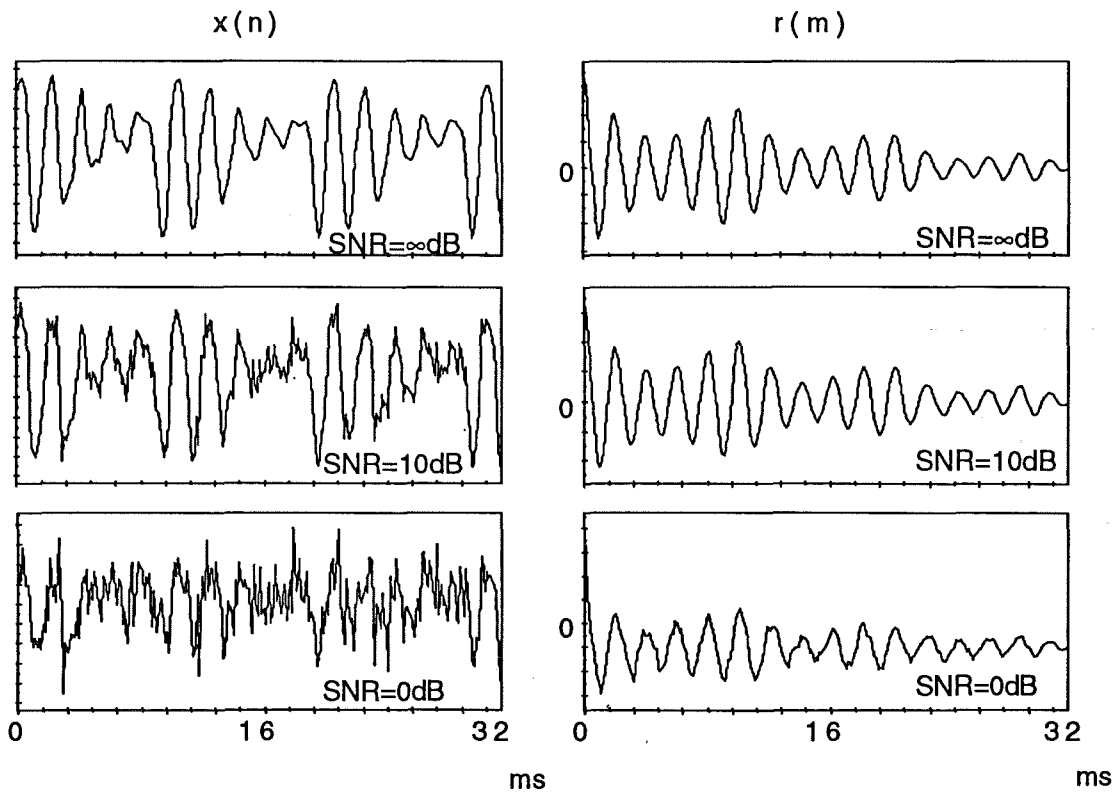


Fig. 3.16. Robustez de la secuencia de autocorrelación

La aplicación de las técnicas de predicción lineal sobre la secuencia de autocorrelación, en lugar de hacerlo sobre la propia señal, puede dar buenos resultados en la estimación de un proceso autorregresivo en presencia de ruido de banda ancha debido a que la secuencia de autocorrelación es más robusta a este tipo de de ruido que la señal.

En la figura 3.16, se muestra un trama de voz sonora sin ruido y contaminada con ruido blanco con una relación señal-ruido de 10 y 0 dB y a su derecha la secuencia de autocorrelación correspondiente calculada utilizando el estimador sesgado clásico. Puede observarse que, aunque el ruido no afecta únicamente al valor de la autocorrelación en el origen, debido a la no idealidad del ruido y a los errores de estimación, la secuencia de autocorrelación es mucho más robusta que la señal.

3.5. PREDICCION LINEAL DE LA PARTE CAUSAL DE LA AUTOCORRELACION

En este apartado se propondrá la predicción lineal de la parte causal de la secuencia de autocorrelación de la señal para la parametrización robusta del habla en presencia de ruido. Esta técnica, que se denotará abreviadamente como OSALPC (*One-Sided Autocorrelation Linear Predictive Coding*), está estrechamente relacionada con las ecuaciones OYWE y OHOYWE, revisadas en el apartado 3.3, y con la técnica de Coherencia Modificada Localizada (SMC, *Short-Time Modified Coherence*), propuesta por Mansour y Juang [Man89a], como se verá más adelante. Su uso en reconocimiento de habla ruidosa es muy interesante debido a su simplicidad, su eficiencia computacional y sus altas tasas de acierto, como se verá en los resultados experimentales presentados en el capítulo 6 de esta memoria.

En el apartado 3.5.1 se presentarán las propiedades de la parte causal de la secuencia de autocorrelación, el espectro analítico y el envolvente espectral. Se verá que existe una correspondencia biunívoca entre el espectro y su envolvente y, por tanto, la estimación de la envolvente del espectro se corresponde con una única estimación del espectro y no representa ninguna pérdida de información. Seguidamente, en el apartado 3.5.2 se presentará una primera técnica muy simple y eficiente de estimación de la envolvente espectral, que se denotará en esta memoria con el nombre de MIAC (Modelado Inverso de la Autocorrelación Causal). Para mejorar las prestaciones de esta estimación se recurrirá en el apartado 3.5.3 al uso de un sistema

sobredeterminado de ecuaciones, lo cual da lugar a la técnica OSALPC. Finalmente, el apartado 3.5.4 tratará de su relación con la representación SMC.

3.5.1. LA PARTE CAUSAL DE LA AUTOCORRELACION. ESPECTRO ANALITICO Y ENVOLVENTE ESPECTRAL

A partir de la secuencia de autocorrelación r(m) de una señal real x(n), se define su parte causal como

$$r^+(m) = \begin{cases} r(m) & m>0 \\ r(0)/2 & m=0 \\ 0 & m<0 \end{cases}, \qquad (3.57)$$

que verifica

$$r^+(m) + r^+(-m) = r(m), \quad -\infty \le m \le \infty. \qquad (3.58)$$

Las transformadas Z y de Fourier de $r^+(n)$, introducidas en análisis espectral por Cadzow [Cad80], se denotarán como $R^+(z)$ y $S^+(\omega)$, respectivamente, es decir,

$$S^+(\omega) = R^+(z)\Big|_{z=e^{j\omega}} = R^+(e^{j\omega}), \qquad (3.59)$$

mientras que las transformadas Z y de Fourier de r(n) se denotarán como R(z) y S(ω), respectivamente. Por tanto, el espectro de la señal S(ω) es

$$S(\omega) = R(z)\Big|_{z=e^{j\omega}} = R(e^{j\omega}). \qquad (3.60)$$

(para simplificar la notación, se prescindirá en adelante del subíndice xx)

Puesto que $r^+(n)$ es una secuencia real y causal y r(m) es dos veces la parte par de $r^+(m)$, se cumple la siguiente relación entre $S^+(\omega)$ y S(ω) [Opp75]

$$S^+(\omega) = \frac{1}{2}\left[S(\omega)+jS_H(\omega)\right], \qquad (3.61)$$

donde $S_H(\omega)$ es la transformada de Hilbert de $S(\omega)$ y responde a la expresión

$$S_H(\omega) = \frac{1}{2\pi} \lim_{\varepsilon \to 0} \left[\int_{\omega+\varepsilon}^{\pi} S(\theta) \cot \frac{\theta-\omega}{2} d\theta + \int_{-\pi}^{\omega+\varepsilon} S(\theta) \cot \frac{\theta-\omega}{2} d\theta \right]. \quad (3.62)$$

Debido a la analogía entre la expresión (3.61) y la definición de señal analítica utilizada en modulación de amplitud, se denominará a $S^+(\omega)$ espectro analítico y a su módulo

$$E(\omega) = |S^+(\omega)| \quad (3.63)$$

envolvente espectral.

Hay una correspondencia biunívoca entre envolvente espectral y espectro. Por tanto, la envolvente espectral no representa ninguna pérdida de información con respecto al espectro. Dado un espectro $S(\omega)$, la envolvente asociada $E(\omega)$ viene dada por las expresiones (3.61)- (3.63). Por otro lado, dada una envolvente espectral $E(\omega)$, el espectro asociado $S(\omega)$ viene definido unívocamente por la expresión

$$S(\omega) = 2 \text{ Re } [S^+(\omega)] = 2 E(\omega) \cos (\Theta(\omega)), \quad (3.64)$$

donde $\Theta(\omega)$ es la curva de fase mínima asociada al módulo $E(\omega)$. Ello es debido que $R^+(z)$ no tiene ceros ni polos fuera de la circunferencia de radio unidad, como se demuestra a continuación.

Si el espectro $S(\omega)$ está acotado, la parte causal de la secuencia de autocorrelación $r^+(n)$ será una secuencia estable y, por tanto, $R^+(z)$ presentará todos sus polos (en caso de tenerlos) en el interior de la circunferencia de radio unidad.

Si, además, el espectro es distinto de cero para cualquier frecuencia, es fácil demostrar que todos los ceros de $R^+(z)$ (en caso de tenerlos) están también en el interior de la circunferencia de radio unidad. Como $r(n)$ es dos veces las parte par de $r^+(n)$, se cumple que

$$\text{Re } [R^+(z)] = \frac{1}{2} R(z). \quad (3.65)$$

Por tanto, al ser el espectro $S(\omega)$ positivo,

$$S(\omega) = R(z)\Big|_{z=e^{j\omega}} = 2\,\mathrm{Re}\left[R^{+}(z)\Big|_{z=e^{j\omega}}\right] > 0 \qquad \text{para todo } \omega \qquad (3.66)$$

Como consecuencia, no existen ceros de $R^{+}(z)$ en la circunferencia de radio unidad. Para demostrar que no existen ceros fuera de esta circunferencia, será suficiente demostrar que

$$\mathrm{Re}\left[R^{+}(z)\Big|_{z=\rho e^{j\omega}}\right] = \mathrm{Re}\left[\sum_{m=0}^{\infty} r^{+}(m)\rho^{-m}e^{-j\omega m}\right] > 0 \qquad (3.67)$$

para todo ω y $1 < \rho \leq \infty$.

Para verificar (3.67) se construye la secuencia $r'(m) = r(m)\, a^{|m|}$, con $0 < a < 1$. La transformada de Fourier de $r'(m)$ será positiva para cualquier frecuencia al ser, salvo un factor de escala, la convolución de $S(\omega)$ con la transformada de Fourier de $a^{|m|}$, la cual es positiva por ser $a^{|m|}$ la autocorrelación de un proceso AR paso-bajo de orden uno. Puesto que la parte causal de $r'(m)$ es $r^{+}(m)\, a^{m}$, se tiene, usando (3.65) que

$$\mathrm{Re}\left[\sum_{m=0}^{\infty} r^{+}(m)a^{m}e^{-j\omega m}\right] = \frac{1}{2}\sum_{m=-\infty}^{\infty} r'(m)e^{-j\omega m} > 0, \qquad (3.68)$$

con lo que queda demostrada (3.67) para todo w y $1 < \rho < \infty$ considerando $\rho = a^{-1}$. En el infinito, el teorema del valor inicial nos garantiza la no existencia de ceros: como $r^{+}(m) = 0$, para $m < 0$, se cumple

$$\lim_{z \to \infty} R^{+}(z) = r^{+}(0) = \frac{r(0)}{2} > 0. \qquad (3.69)$$

Por tanto, si el espectro está acotado y es diferente de cero para cualquier frecuencia, entonces $R^{+}(z)$ tiene los polos y ceros en el interior de la circunferencia unidad, es decir, $r^{+}(m)$ es una secuencia de fase mínima.

Si el espectro está acotado y es igual a cero para algunas frecuencias, $R^{+}(z)$ tiene los polos y ceros en el interior de la circunferencia unidad a menos que exista

simetría par del espectro respecto a alguno de sus ceros. Ello es debido a que la función $\cot((\theta+\omega)/2)$, que aparece en la relación de Hilbert (3.62), es una función par en θ. Como consecuencia, para que siendo $S(\omega_0) = 0$ se cumpla $S_H(\omega_0) = 0$ debe verificarse que $S(\theta+\omega_0)$ sea una función par en θ; es decir, $S(\omega)$ ha de presentar simetría par con respecto a ω_0. Esta situación, a menos que el espectro presente un cero en $\omega = 0$, es bastante inusual, por lo que $r^+(m)$ es una secuencia de fase mínima en la mayoría de los casos. Si el espectro presenta simetría par respecto a algunos de sus ceros, $R^+(z)$ tiene todos los polos y los ceros en el interior de la circunferencia unidad a excepción de los ceros que verifiquen tal condición, que se encuentran en dicha circunferencia.

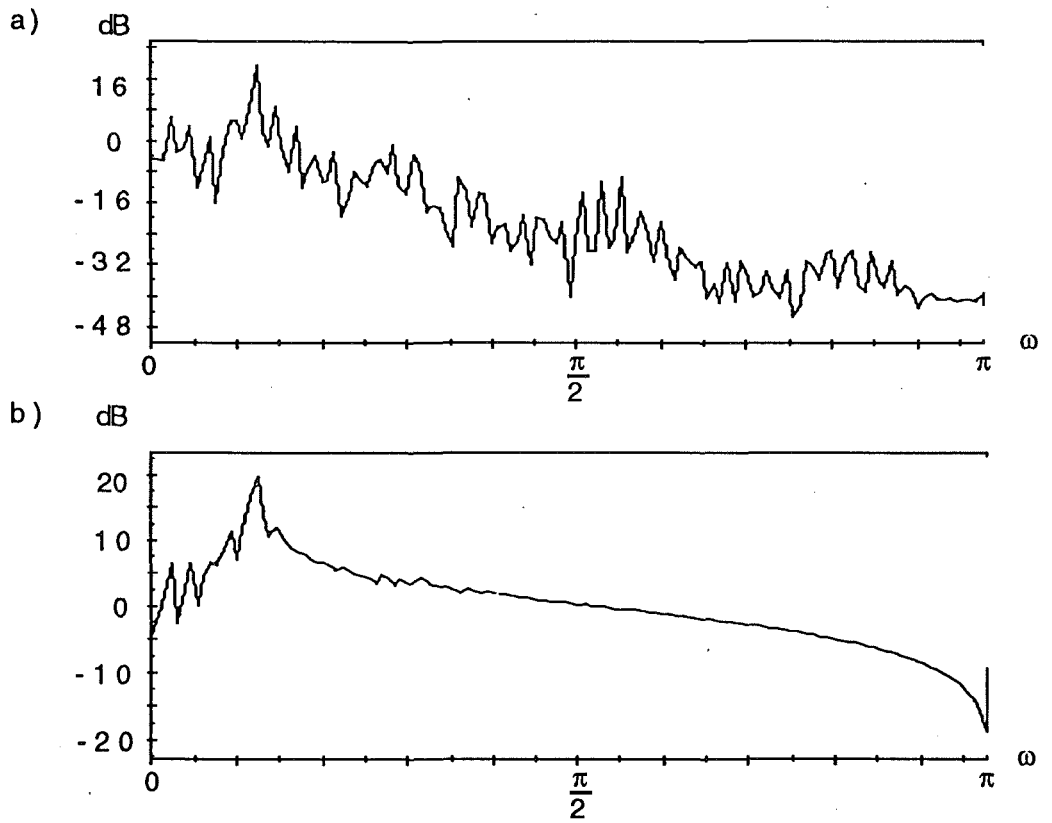


Fig. 3.17. Espectro (periodograma) (a) y envolvente espectral (b) de una trama de voz sonora. Vocal /o/ del dígito catalán "dos"

En cualquier caso, $R^+(z)$ nunca tiene polos ni ceros fuera de la circunferencia de radio unidad y, por tanto, existe una correspondencia biunívoca entre espectro y

envolvente espectral. Teniendo en cuenta que la envolvente espectral no representa ninguna pérdida de información con respecto al espectro, es una buena candidata para ser utilizada en estimación espectral [Ame88] [Nad89].

De la expresión (3.62) se deduce que la envolvente espectral $E(\omega)$ es más suave que el espectro asociado $S(\omega)$, ya que él término $\cos(\Theta(\omega))$ introduce variaciones en $S(\omega)$ no existentes en $E(\omega)$. Esto puede observarse claramente en la figura 3.17.

Este carácter de envolvente, junto con el alto rango dinámico del espectro de voz, origina que $E(\omega)$ enfatice las bandas de frecuencia de mayor potencia, que son precisamente las más robustas a un ruido de banda ancha. Por tanto, $E(\omega)$ es más robusta a este tipo de ruido que $S(\omega)$.

Teniendo en cuenta que el cuadrado de la envolvente espectral $E^2(\omega)$ es precisamente el espectro de $r^+(m)$ (elevar al cuadrado ambos términos de la expresión (3.63)), el párrafo anterior es equivalente a afirmar que el espectro de $r^+(m)$, $E^2(\omega)$, es más robusto a este tipo de ruido que el espectro de la propia señal $x(n)$, $S(\omega)$. Esta propiedad también puede constatarse en el dominio temporal en la figura 3.16.

Por otro lado, suponiendo un modelo autorregresivo de la señal de voz, los polos de la transformada Z de la parte causal de su autocorrelación $r^+(m)$, $R^+(z)$, son los mismos que los de la transformada Z de la propia señal de voz $x(n)$, $X(z)$, como se verá en el apartado siguiente.

Ambos factores sugieren que los parámetros autorregresivos de la señal de voz pueden ser estimados de forma más fiable aplicando las técnicas de predicción lineal clásicas vistas en el apartado 3.2 sobre $r^+(m)$, en lugar de sobre la propia señal $x(n)$, cuando la señal de voz está contaminado por ruido de banda ancha. Esta es la base de la técnica OSALPC, que se presentará en el apartado 3.5.3.

No obstante, antes de presentar esta técnica, en el apartado 3.5.2 se describirá el método MIAC, que permite estimar de un modo muy simple y eficiente los parámetros autorregresivos de la señal de voz realizando un modelado todo-polos de $R^+(z)$ y, por tanto, de $E^2(\omega)$. Este método, aunque no destaca por sus prestaciones en reconocimiento robusto del habla, permitirá introducir de una manera simple la técnica OSALPC, cuya utilización en reconocimiento de habla ruidosa, como ya se ha

comentado, es muy interesante debido a su simplicidad, su eficiencia computacional y sus altas tasas de acierto

Por último, es importante hacer notar que la envolvente espectral es una función de cuarto orden y este trabajo podría haberse enmarcado en el estudio de la utilización de momentos de orden superior en estimación espectral. Sin embargo, el concepto de envolvente resulta por sí mismo explicativo y, por ello, se han concentrado los esfuerzos en esta función especial de cuarto orden sin generalizar a otras funciones.

3.5.2. MODELADO INVERSO DE LA AUTOCORRELACION CAUSAL (MIAC)

Si x(n) es un proceso real autorregresivo de orden p, cuyo espectro viene dado por la expresión

$$S(\omega) = \frac{G^2}{|A(e^{j\omega})|^2}, \qquad (3.70)$$

con

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \qquad (3.71)$$

la transformada Z de su autocorrelación R(z) será

$$R(z) = \frac{G^2}{A(z)A(z^{-1})}. \qquad (3.72)$$

Como A(z) es el denominador de la función de transferencia del filtro del modelo H(z), que se supone causal y estable, los ceros de A(z) estarán en el interior de la circunferencia de radio unidad. Hay que destacar, no obstante, que entre todas las técnicas de estimación AR mencionadas hasta ahora sólo garantiza estabilidad el método de autocorrelación de predicción lineal.

Por otro lado, es fácil comprobar que R(z) puede escribirse en función de la transformada Z de la parte causal de la secuencia de autocorrelación $R^+(z)$ como

$$R(z) = R^{+}(z) + R^{+}(z^{-1}) . \qquad (3.73)$$

Por tanto, podrá escribirse (3.72) de la forma

$$R(z) = \frac{C(z)}{A(z)} + \frac{C(z^{-1})}{A(z^{-1})} , \qquad (3.74)$$

donde el primer término se corresponderá con $R^{+}(z)$ por tener los polos en el interior de la circunferencia unidad (ver apartado anterior), es decir,

$$R^{+}(z) = \frac{C(z)}{A(z)} . \qquad (3.75)$$

Como conclusión, la transformada Z de la parte causal de la autocorrelación $R^{+}(z)$ tiene los mismos polos que el filtro H(z) y, por tanto, los mismos polos que señal [McG83].

En cuanto a los ceros de $R^{+}(z)$, C(z) será un polinomio en z^{-1} por ser $r^{+}(m)$ una secuencia causal. Por otro lado, a partir de (3.71) y (3.75) se obtiene que $C(\infty)$ es igual a $R^{+}(\infty)$, que aplicando el teorema del valor inicial coincide con $r^{+}(0)$. Como consecuencia, el término independiente de C(z) es $r^{+}(0)$ y, por tanto, no nulo. Teniendo en cuenta que el término independiente de C(z) es no nulo en la expresión

$$G^2 = C(z)A(z^{-1}) + C(z^{-1})A(z), \qquad (3.76)$$

que resulta de identificar (3.72) y (3.74), se obtiene fácilmente que C(z) es un polinomio del mismo orden que A(z).

Además, como el espectro (3.70) es diferente de cero en la práctica para cualquier frecuencia, $r^{+}(m)$ es una secuencia de fase mínima y, por tanto, los p ceros de A(z), que coinciden con los polos de la señal, y los p ceros de C(z) están en el interior de la circunferencia de radio unidad.

Finalmente, $R^{+}(z)$ puede escribirse como

$$R^+(z) = \frac{c_0 + \sum_{k=1}^{p} c_k z^{-k}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad (3.77)$$

y las expresiones para el espectro analítico y el cuadrado de la envolvente espectral serán

$$S^+(\omega) = \frac{C(e^{j\omega})}{A(e^{j\omega})} \qquad (3.78)$$

y

$$E^2(\omega) = \frac{|C(e^{j\omega})|^2}{|A(e^{j\omega})|^2}, \qquad (3.79)$$

respectivamente.

Hay que hacer notar que, aunque aparezcan 2p+1 parámetros en la expresión (3.77), sólo p+1 son independientes, ya que R(z) queda especificada por G^2 y p coeficientes de A(z) y esta tiene una correspondencia biunívoca con $R^+(z)$. La relación de dependencia es (3.76). A partir de ella, puede calcularse C(z) a partir de G^2 y A(z).

En el dominio temporal, la expresión (3.77) se convierte en

$$r^+(m) = - \sum_{k=1}^{p} a_k \, r^+(m-k) + \sum_{k=0}^{p} c_k \, \delta(m-k), \qquad (3.80)$$

donde $\delta(n)$ es el impulso unidad. Esta expresión se convierte en identidad para $m < 0$, pues todos los términos son nulos. Por tanto, sólo se considerará para $m \geq 0$.

Una posible forma de hallar los parámetros del modelo a partir de la señal de voz es evaluar la expresión (3.80) para m = 0,... 2p y resolver el sistema de ecuaciones resultante (3.81) utilizando un estimador adecuado de los valores de r(m) y usando la relación (3.57) entre $r^+(m)$ y r(m).

$$\begin{pmatrix} r(0)/2 & 0 & 0 & \dots & 0 \\ r(1) & r(0)/2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0)/2 \\ r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ \vdots & \vdots & \vdots & & \vdots \\ r(2p) & r(2p-1) & r(2p-2) & \dots & r(p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (3.81)$$

La primera ecuación constata el hecho antes mencionado de que el término independiente c_0 de C(z) es $r^+(0)=r(0)/2$. Por otro lado, las últimas p ecuaciones también resultan desacopladas del resto y nos proporcionan la estimación de los coeficientes a_k. Sin embargo, este subsistema de ecuaciones es precisamente el mismo que el de las ecuaciones de Yule-Walker de orden superior (HOYWE), que se han descrito en el apartado 3.3.2.

En este trabajo se propone modelar $R^+(z)$ como una función todo polos, aprovechando el hecho ya mencionado de que un modelo todo-polos permite aproximar cualquier modelo racional utilizando un número suficientemente elevado de polos. Esto equivale a incrementar el valor de p y suponer que todos los ceros de C(z) están en el origen y, por tanto,

$$C(z) = c_0 = r^+(0). \qquad (3.82)$$

En este caso, las expresiones (3.77) y (3.80) pasan a ser

$$R^+(z) = \frac{r^+(0)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad (3.83)$$

$$r^+(m) = - \sum_{k=1}^{p} a_k \, r^+(m-k) + r^+(0) \quad \delta(m). \qquad (3.84)$$

La expresión (3.84) se convierte en identidad para $m < 0$, pues todos los términos son nulos, y para $m = 0$, pues ambos miembros de la igualdad son $r^+(0)$. Por tanto, sólo es necesario considerarla para $m > 0$.

Una posible forma de estimar los coeficientes a_k a partir de la señal de voz es evaluar (3.84) para $m = 1,\dots p$ y resolver el sistema de ecuaciones resultante (3.85)

utilizando un estimador adecuado de los valores de r(m) y usando la relación (3.57) entre $r^+(m)$ y r(m).

$$\begin{pmatrix} r(1) & r(0)/2 & 0 & \dots & 0 \\ r(2) & r(1) & r(0)/2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0)/2 \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.85}$$

Este método de estimación de los coeficientes a_k se denotará en esta memoria como MIAC, Modelado Inverso de la Autocorrelación Causal, debido al modelo $R^+(z)$ en que está basado (3.83).

El sistema de ecuaciones (3.85) también puede escribirse como

$$\begin{pmatrix} r(0)/2 & 0 & \dots & 0 \\ r(1) & r(0)/2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & r(0)/2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = - \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{pmatrix} \tag{3.86}$$

Al tratarse de un sistema triangular su resolución es muy simple. Su coste computacional es mucho menor que el del algoritmo de Levinson-Durbin que se utiliza para resolver las ecuaciones de Yule-Walker (YWE), el más eficiente y popular de los métodos de predicción lineal.

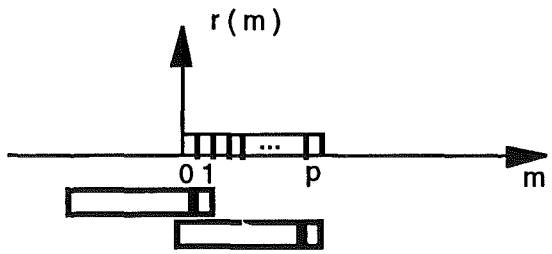


Fig. 3.18. Representación esquemática del método MIAC

La representación esquemática del método MIAC puede verse en la figura (3.18). De acuerdo con esta figura, el método MIAC puede interpretarse como la predicción lineal exacta de orden p de los valores de la secuencia $r^+(m)$ desde m = 1 a p utilizando el método de covarianza, ya que no supone ningún enventanado de dicha secuencia.

También puede interpretarse el método MIAC como la predicción lineal exacta de orden p de los valores de la secuencia de autocorrelación r(m), sustituyendo r(0) por r(0)/2, desde m = 1 a p utilizando el método de preenventanado, ya que se suponen nulos los valores de la secuencia de autocorrelación anteriores al intervalo de predicción.

Recuérdese que en el apartado 3.4. se interpretó el uso de las ecuaciones de Yule-Walker como la predicción lineal exacta de orden p de los valores de la secuencia de autocorrelación r(m) desde m = 1 a p utilizando el método de covarianza (ver figura 3.12).

El método MIAC es muy simple y eficiente, pero no destaca por sus prestaciones en reconocimiento robusto del habla. Sin embargo, permite introducir de una manera simple la técnica OSALPC.

3.5.3. PREDICCION LINEAL DE LA PARTE CAUSAL DE LA AUTOCORRELACION (OSALPC)

Una aproximación alternativa al método MIAC para mejorar la estimación de los coeficientes a_k del modelo todo-polos de $R^+(z)$ (3.83) es el uso de un sistema sobredeterminado de más de p ecuaciones obtenidas evaluando (3.84) para valores de m mayores que 0.

Una justificación del uso de un sistema de ecuaciones sobredeterminado es que en las ecuaciones (3.85) sólo intervienen los valores estimados de la autocorrelación de m = 0 a p y, por tanto, los coeficientes a_k obtenidos dependen totalmente de las estimaciones de la secuencia de autocorrelación para estos valores de m, que presentan un cierto error de estimación. Además, hay que tener en cuenta el efecto de bordes. Estos errores de estimación pueden compensarse mediante la utilización de más del mínimo número p de ecuaciones, con lo cual se hace intervenir en la obtención de los coeficientes a_k un conjunto mayor de valores estimados de autocorrelación. Recordar que estos mismos motivos conducen a la propuesta de las OHOYWE a partir de la HOYWE.

Por otro lado, si se pretende realizar una estimación fiable de los coeficientes a_k en presencia de ruido, la utilización de valores de autocorrelación alejados del origen puede ser favorable pues estos son más robustos a un ruido de espectro plano que los cercanos al origen.

Suponiendo que M, tal que M > p, es el mayor índice para el que la autocorrelación puede estimarse con cierta fiabilidad, puede construirse el siguiente sistema sobredeterminado de M ecuaciones y p incógnitas evaluando (3.84) desde m = 1 hasta M

$$\begin{pmatrix} r(1) & r(0)/2 & 0 & \dots & 0 \\ r(2) & r(1) & r(0)/2 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0)/2 \\ r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(p) \\ \varepsilon(p+1) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix}, \qquad (3.87)$$

donde $\varepsilon(m)$ es el error asociado a la estimación de las autocorrelaciones. Este sistema de ecuaciones puede resolverse minimizando el error cuadrático y utilizarse como un método fiable de estimación de los coeficientes a_k del modelo todo-polos de $R^+(z)$ (3.83) utilizando un estimador adecuado de los valores de la autocorrelación.

El sistema de ecuaciones (3.87) es equivalente al sistema

$$\begin{pmatrix} r(0)/2 & 0 & 0 & \dots & 0 \\ r(1) & r(0)/2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0)/2 \\ r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(0) \\ \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(p) \\ \varepsilon(p+1) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix}, \qquad (3.88)$$

ya que simplemente se ha añadido la ecuación $r(0)/2 = \varepsilon(0)$, la cual no depende de la incógnitas a_k.

Comparando el sistema de ecuaciones (3.88) con el sistema (3.13), si se toma como matiz de datos $\mathbf{X}_3$, es claro este método equivale a realizar predicción lineal de orden p de la secuencia de autocorrelación, sustituyendo r(0) por r(0)/2, mediante el método de preenventanado. En efecto, el sistema de ecuaciones (3.88) es idéntico al sistema (3.13), usado en predicción lineal de la señal, considerando una secuencia de autocorrelación r(m) de m = 0 a M (cambiando r(0) por r(0)/2), en lugar de una secuencia de señal x(n) de n = 1 a N, y sustituyendo el error de predicción de la señal e(n) por el error de predicción de la autocorrelación $\varepsilon(m)$.

Las pruebas experimentales en reconocimiento del habla ruidosa realizadas en este trabajo (ver capítulo 6) muestran que este método de estimación de los coeficientes a_k, utilizando el estimador sesgado de la autocorrelación, proporciona tasas de reconocimiento en condiciones severas de ruido notablemente superiores a los métodos descritos en los apartados anteriores.

Debido al carácter Toeplitz de la matriz de coeficientes del sistema (3.88), éste puede resolverse de forma eficiente [Fri79]. Sin embargo, el método de autocorrelación de predicción lineal es más eficiente gracias a la posibilidad de aplicar el algoritmo de Levinson-Durbin. Por tanto, resultaría atractiva la aplicación del método de autocorrelación sobre r(m), en lugar del método de preenventanado, si no se produjera un merma en las prestaciones. Esto es plausible ya que la secuencia de autocorrelación, calculada utilizando el estimador sesgado clásico, tiene un carácter globalmente decreciente y, por tanto, los efectos del enventanado hasta un índice M suficientemente grande pueden no ser importantes.

Las pruebas experimentales presentadas en el capítulo 6 de esta memoria muestran que la aplicación del método de autocorrelación de predicción lineal sobre la secuencia de autocorrelación r(m) de m = 0 a M (sustituyendo r(0) por r(0)/2), equivalente a encontrar los coeficientes a_k que minimizan el error cuadrático en el sistema

$$\begin{pmatrix} r(0)/2 & 0 & 0 & \dots & 0 \\ r(1) & r(0)/2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0)/2 \\ r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \\ 0 & r(M) & r(M-1) & \dots & r(M-p+1) \\ 0 & 0 & r(M) & \dots & r(M-p+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & r(M) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(0) \\ \varepsilon(1) \\ \vdots \\ \varepsilon(p) \\ \varepsilon(p+1) \\ \vdots \\ \vdots \\ \varepsilon(M) \\ \varepsilon(M+1) \\ \varepsilon(M+2) \\ \vdots \\ \varepsilon(M+p) \end{pmatrix} \qquad (3.89)$$

tiene prestaciones tan notables en reconocimiento de habla ruidosa como la aplicación del método de preenventanado, correspondiente al sistema de ecuaciones (3.88). Como consecuencia, el método de autocorrelación resulta más atractivo que el de preenventanado.

Esta nuevo método de estimación de los coeficientes a_k se denota en esta memoria con las siglas OSALPC (*One-Sided Autocorrelation Linear Predictive Coding*) y su utilización en reconocimiento de habla ruidosa resulta muy interesante debido a su simplicidad, su eficiencia computacional y sus altas tasas de acierto [Her92d].

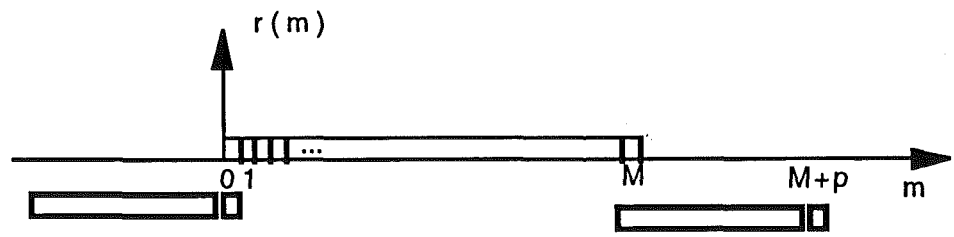


Fig. 3.19. Representación esquemática de la técnica OSALPC

La representación esquemática de la técnica OSALPC puede verse en la figura (3.19). Como ya se ha mencionado, esta técnica puede interpretarse como la predicción lineal de orden p de los valores de la secuencia de autocorrelación r(m), sustituyendo r(0) por r(0)/2, desde m = 1 a M utilizando el método de autocorrelación, ya que se suponen nulos los valores de la secuencia de autocorrelación anteriores y posteriores

al intervalo de predicción. Sin embargo, de acuerdo con la figura (3.19), también puede interpretarse como la predicción lineal de orden p de los valores de la secuencia $r^+(m)$ desde m = 1 a M utilizando el método de postenventanado, ya que sólo se suponen nulos los valores de dicha secuencia posteriores al intervalo de predicción.

En ambas interpretaciones se supone un enventanado de la secuencia hasta un índice M suficientemente grande. Sin embargo, sólo en la primera interpretación es necesario suponer valores nulos cercanos al origen, cuyos efectos en la estimación pueden ser importantes. Por tanto, la segunda interpretación de la técnica OSALPC como predicción lineal de la parte causal de la secuencia de autocorrelación es más realista. De ahí el nombre dado a la técnica OSALPC (en español, predicción lineal de la parte causal de la autocorrelación).

Su implementación práctica es muy simple. Una vez estimados los valores de la parte causal de la secuencia de autocorrelación $r^+(m)$ desde m = 0 a M, se aplica sobre dicha secuencia el método de la autocorrelación de predicción lineal. Para ello, se calculan los coeficientes r'(m) de las ecuaciones de Yule-Walker (3.20) aplicando el estimador sesgado de la autocorrelación sobre la parte causal de la secuencia de autocorrelación, es decir,

$$r'(m) = \sum_{n=0}^{M-m} r^+(n+m)\, r^+(n), \qquad (3.90)$$

y, finalmente, se resuelven dichas ecuaciones utilizando el algoritmo de Levinson-Durbin.

Por tanto, una vez estimada la parte causal de la secuencia de autocorrelación $r^+(m)$ desde m = 0 a M, el coste computacional de esta técnica es el mismo que el del método de autocorrelación de predicción lineal aplicado sobre una trama de M+1 puntos. El mayor esfuerzo de cálculo lo supone el cálculo de las autocorrelaciones.

Interpretando la técnica OSALPC como predicción lineal de la parte causal de la secuencia de autocorrelación, esta técnica se corresponde con un modelado todo-polos del espectro de $r^+(m)$, el cuadrado de la envolvente espectral, es decir

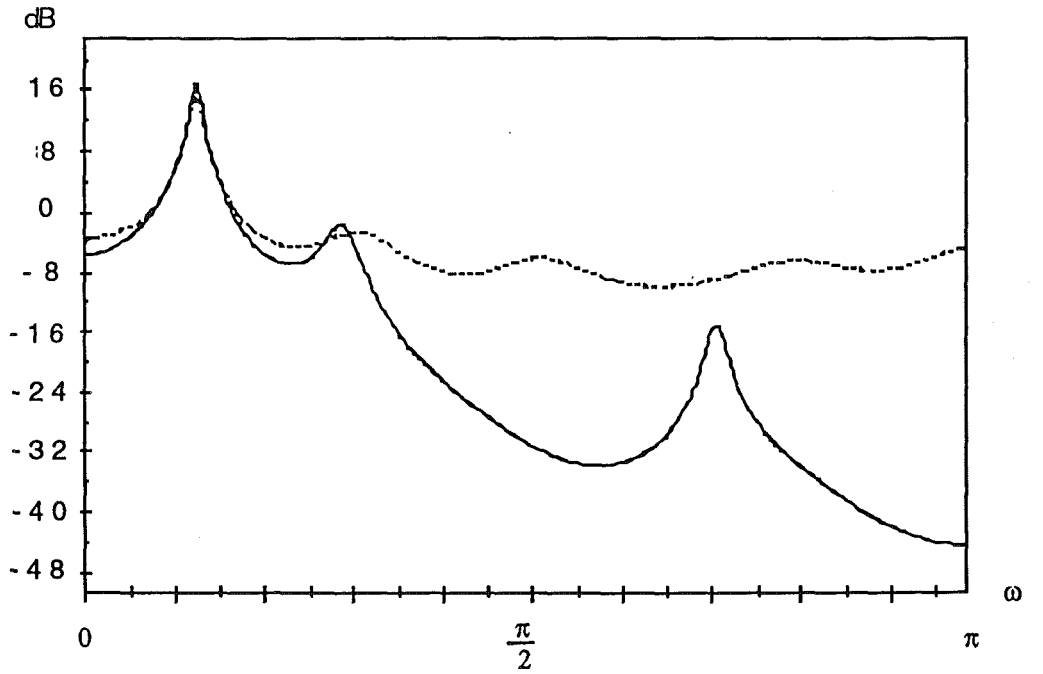
$$E^2(\omega) = \frac{(r^+(0))^2}{|A(e^{j\omega})|^2}, \qquad (3.91)$$

donde el polinomio A(z) del denominador es (3.71) y el numerador se obtiene directamente a partir del modelo de $R^+(z)$ (3.83).

También puede llegarse al mismo resultado desde el punto de vista de modelado paramétrico de procesos. Puede observarse en la expresión (3.90) que los coeficientes de las ecuaciones de Yule-Walker son una estimación de la autocorrelación de la parte causal de la secuencia de autocorrelación. Por tanto, la técnica OSALPC supone un modelado un modelado AR de la parte causal de la secuencia de autocorrelación y, por tanto, su espectro $E^2(\omega)$ es todo-polos.

En la figura 3.20. se comparan las estimaciones espectrales correspondientes al método de autocorrelación de predicción lineal sobre la señal (a), abreviadamente en la figura LPC, y a la técnica OSALPC (b) para p = 12, en condiciones supuestamente libres de ruido (línea continua) y en presencia de ruido blanco gaussiano aditivo con un relación señal-ruido de 0 dB (línea de puntos). No se ha tenido en cuenta en la representación el término de ganancia del numerador de ambos modelos.

a) Espectro LPC



(b) Cuadrado de la envolvente OSALPC

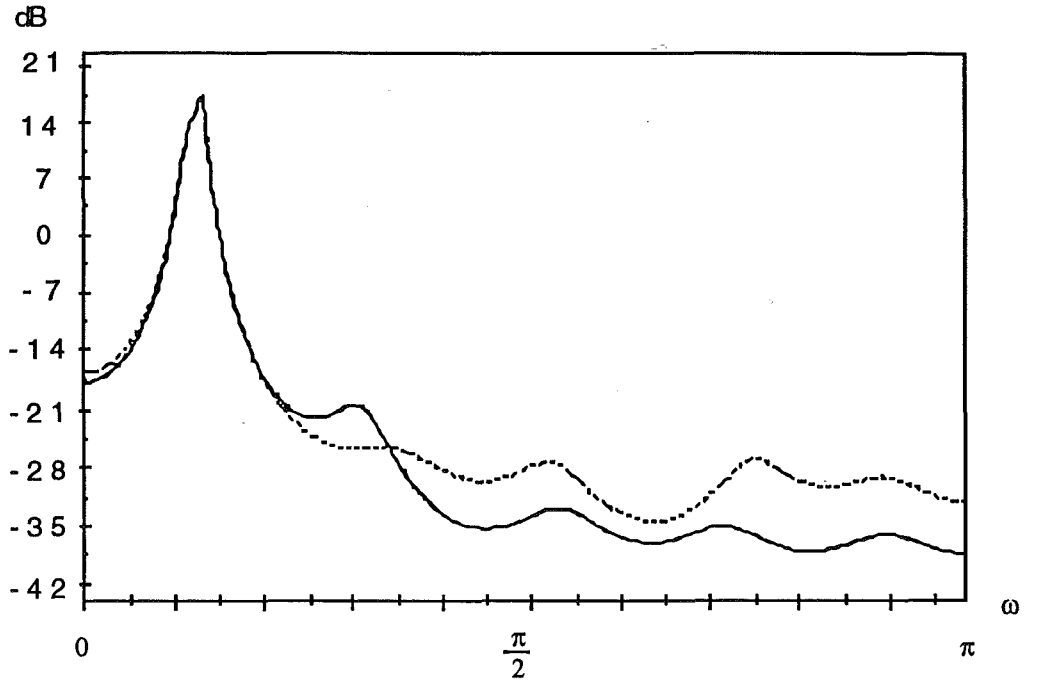


Fig. 3.20 . Robustez de la técnica OSALPC al ruido blanco aditivo: (a) espectro LPC y (b) cuadrado de la envolvente OSALPC de una trama de la parte estacionaria de la vocal /o/ en condiciones libres de ruido (línea continua) y SNR = 0 dB (línea discontinua).

En cuanto a la robustez de ambas técnicas frente al ruido, puede observarse en la figura que en ambos casos el primer formante no es prácticamente alterado por el ruido. Sin embargo, en el resto de las frecuencias hay una clara diferencia entre la robustez frente al ruido de las dos técnicas. El espectro correspondientes a la técnica clásica es muy sensible a la presencia del ruido: se produce una espectacular reducción del rango dinámico y la estructura de formantes a partir del segundo formante queda totalmente alterada, incluso aparece un nuevo formante. Sin embargo, el cuadrado de la envolvente espectral OSALPC es mucho más robusto al ruido: se mantiene el margen dinámico y sólo cambian ligeramente la frecuencia central y el ancho de banda de los formantes siguientes al primero. Por tanto, queda claro que la técnica OSALPC es mucho más robusta al ruido que la técnica clásica y puede esperarse que sus prestaciones en reconocimiento de habla ruidosa sean superiores, si se conserva la capacidad discriminativa entre diferentes sonidos.

También puede observarse en la figura que la envolvente espectral enfatiza fuertemente las bandas de frecuencia de mayor potencia, hecho ya comentado en el apartado 3.5.1.

Por otro lado, si se comparan las dos estimaciones espectrales en el caso de ausencia de ruido, se observa que en la técnica OSALPC aparecen formantes espurios con respecto a la técnica de predicción clásica. Este hecho puede explicarse teniendo en cuenta que el modelado espectral asociado a la técnica de predicción lineal clásica, consistente con el modelo lineal de producción de voz, equivale a un modelado autorregresivo de la señal de voz. Considerando la señal de voz un proceso autorregresivo, el cuadrado de la envolvente espectral de la señal de voz es una función con polos y ceros (3.79), como se ha demostrado en el apartado anterior. Sin embargo, la técnica OSALPC se ha derivado a partir de la simplificación (3.82), que supone un modelo todo-polos para el cuadrado de la envolvente espectral (3.91).

Por último, es importante destacar que escogiendo diferente conjunto de ecuaciones a partir del sistema (3.89) que define la técnica OSALPC surgen otros métodos de estimación espectral ya vistos anteriormente. En la figura 3.21 puede observarse la relación de la técnica OSALPC con el resto de las técnicas vistas hasta el momento. La calidad de las estimaciones espectrales proporcionadas por cada una de estas técnicas depende del compromiso robustez al ruido-varianza en función del índice m de las estimaciones de los valores de la secuencia de autocorrelación r(m), ya comentado, y del distinto modo de enventanar dicha secuencia.

3.5.4. RELACION CON LA COHERENCIA MODIFICADA LOCALIZADA (SMC)

La técnica OSALPC propuesta en el apartado anterior está también estrechamente relacionada con la representación SMC (*Short-Time Modified Coherence*, Coherencia Modificada Localizada) [Man89a].

D. Mansour y B.H. Juang propusieron el uso de la predicción lineal de la secuencia de autocorrelación para la obtención de una parametrización robusta de la señal de voz, basándose en que la secuencia de autocorrelación de un proceso AR de modelo G/A(z) es también un proceso AR de modelo

$$R(z) = \frac{G^2}{A^2(z)}, \qquad (3.92)$$

$$\begin{pmatrix} r(0)/2 & 0 & 0 & \dots & 0 \\ r(1) & r(0)/2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0)/2 \\ r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ \vdots & \vdots & \vdots & & \vdots \\ r(2p) & r(2p+1) & r(2p+2) & \dots & r(p) \\ r(2p+1) & r(2p+2) & r(2p+3) & \dots & r(p+1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \\ 0 & r(M) & r(M-1) & \dots & r(M-p+1) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & r(M) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(0) \\ \varepsilon(1) \\ \vdots \\ \varepsilon(p) \\ \varepsilon(p+1) \\ \vdots \\ \varepsilon(2p) \\ \varepsilon(2p+1) \\ \vdots \\ \vdots \\ \varepsilon(M) \\ \varepsilon(M+1) \\ \vdots \\ \varepsilon(M+p) \end{pmatrix}$$

(1) (2) (3) (4)

$$\begin{pmatrix} r(1) & r(0) & r(2) & \dots & r(p-1) \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \dots & r(0) \\ r(p+1) & r(p) & r(p-1) & \dots & r(1) \\ \vdots & \vdots & \vdots & & \vdots \\ r(2p) & r(2p+1) & r(2p+2) & \dots & r(p) \\ r(2p+1) & r(2p+2) & r(2p+3) & \dots & r(p+1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M) & r(M-1) & r(M-2) & \dots & r(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(1) \\ \vdots \\ \varepsilon(p) \\ \varepsilon(p+1) \\ \vdots \\ \varepsilon(2p) \\ \varepsilon(2p+1) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix}$$

(5) (2) (3) (6)

(1) MIAC (Modelado Inverso de la Autocorrelación Causal)

(2) HOYWE (*High Order Yule-Walker Equations*)

(3) OHOYWE (*Overdetermined High Order Yule-Walker Equations*)

(4) OSALPC (*One-Sided Autocorrelation Linear Predictive Coding*)

(5) YWE (*Yule-Walker Equations*), LPC (Predicción Lineal Clásica)

(6) OYWE (*High Order Yule-Walker Equations*)

Fig. 3.21. Relación de la técnica OSALPC con otras técnicas de predicción lineal

es decir, tiene los mismos polos que la señal con multiplicidad doble.

De acuerdo con el modelo (3.92), sería necesario aumentar el orden de predicción de p a 2p. Este incremento artificial del orden no sólo aumenta la complejidad del modelo sino que conlleva una dificultad adicional en la práctica cuando el análisis de orden 2p no contiene exactamente p polos dobles. Además, la interacción entre el tono y los formantes de la voz será mucho más pronunciada, especialmente para locutores femeninos, lo cual lleva a una alta variabilidad de los coeficientes de predicción y provoca un empeoramiento de las tasas de reconocimiento.

Para corregir estos efectos, la representación SMC introduce un conformador espectral en forma de raíz cuadrada, que reduce el margen dinámico en un factor 2 en el logaritmo del espectro. De este modo, el orden del modelo es de nuevo p y la interacción entre el tono y los formantes se vuelve esencialmente la misma que en la predicción lineal clásica sobre la señal.

En la figura 3.22 se representa el algoritmo de cálculo de la representación SMC propuesto en [Man89a]. En primer lugar, a partir de la trama de señal x(n), desde n=1 a N, se estiman los valores de la secuencia de autocorrelación r(m) desde m=0 a N/2 utilizando la expresión

$$r(m) \approx \sum_{n=0}^{N/2-1} x(n)x(n+m) \qquad m = 0,\ldots N/2. \qquad (3.93)$$

Este estimador de la autocorrelación recibe el nombre de coherencia, de ahí el nombre del método, y sus propiedades serán abordadas más adelante.

Posteriormente se enventana la secuencia de autocorrelación. Debido a que el margen dinámico de la secuencia de autocorrelación en el dominio frecuencial es el doble que en el caso de la señal, una ventana rectangular podría enmascarar formantes de amplitud baja. Se elige la ventana de Hamming como compromiso entre resolución y valores bajos de lóbulos laterales. Otra razón para aplicar la ventana de Hamming es que sobre esta secuencia se va a realizar predicción lineal y en los extremos se encuentran transitorios demasiado fuertes.

A continuación se realiza la FFT sobre la secuencia de autocorrelación enventanada (el término del origen se anula, ya que no es necesario en general y sí perjudicial cuando la señal está contaminada por ruido blanco) y FFT inversa sobre el módulo del resultado de dicha FFT. De este modo, se obtiene una estimación de la autocorrelación de la secuencia de autocorrelación enventanada anterior, pero

conformada espectralmente para mantener el margen dinámico en el dominio frecuencial y evitar el incremento artificial de orden de predicción.

Finalmente, se aplica el algoritmo de Levison-Durbin utilizando como entradas los valores de m=0 a p de esta secuencia de autocorrelación conformada espectralmente.

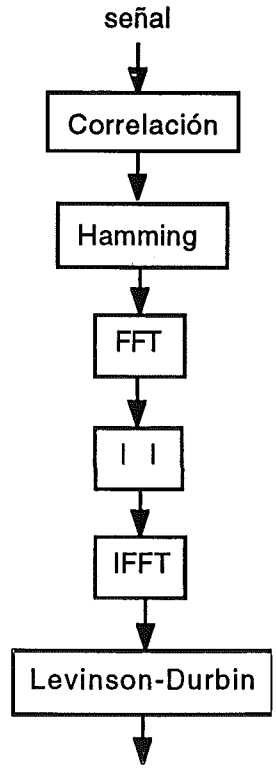


Fig. 3.22. Cálculo de la representación SMC

La diferencia fundamental entre la representación SMC y la técnica OSALPC introducida en el apartado anterior es la utilización del conformador espectral, derivado de la utilización del modelo de autocorrelación $G^2/A^2(z)$ (3.92) para un proceso AR de modelo G/A(z). Este modelo es incorrecto [García Gómez y Gómez Mena]. Observar, por ejemplo, los modelos dados en este capítulo para la autocorrelación (3.72) y la parte causal de la autocorrelación (3.75) de un proceso AR en este capítulo. En el capítulo de resultados experimentales de esta memoria, se observa que la técnica OSALPC supera en prestaciones a la SMC.

En términos de la formulación introducida en este capítulo, la técnica OSALPC realiza un modelado espectral mediante predicción lineal del cuadrado de la envolvente espectral $E^2(\omega)$, mientras la representación SMC realizaría un modelado espectral mediante predicción lineal de la envolvente misma $E(\omega)$.

Por último, es importante mencionar las propiedades del estimador coherencia utilizado por la técnica SMC para obtener la secuencia de autocorrelación, sobre la que posteriormente se aplica predicción lineal.

En la figura 3.23 están esquematizadas las operaciones que se realizan para obtener cada valor de autocorrelación r(m). Como puede observarse en dicha figura, para referirse a este estimador de autocorrelación es más propio hablar de coherencia entre dos segmentos adyacentes de señal, que corresponden a las dos mitades de la trama.

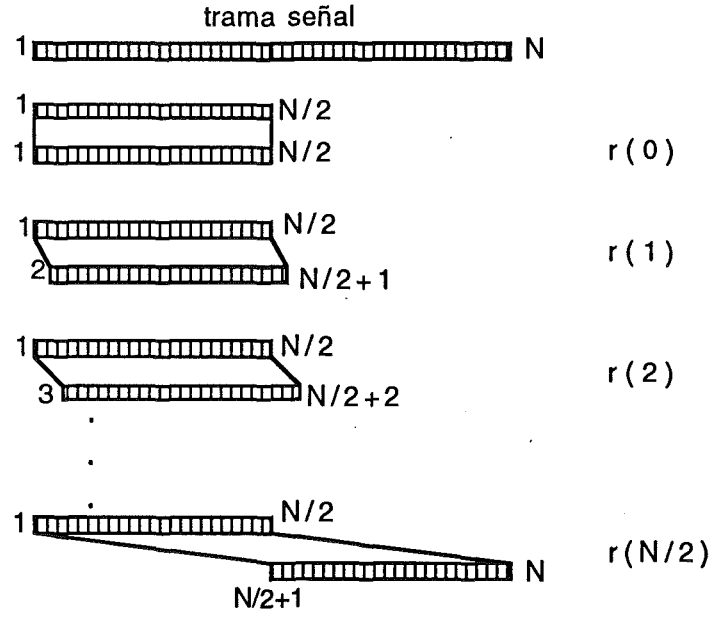


Fig. 3.23. Diagrama de cálculo de la coherencia

En particular, el valor estimado de la autocorrelación en el origen coincide con el valor que se obtendría aplicando el estimador sesgado clásico sobre la primera mitad de la trama y el valor estimado para M=N/2 se corresponde con el valor en el origen de la correlación cruzada entre las dos mitades de la trama. Los valores intermedios

proporcionados por este estimador proporciona una medida de similitud entre ambas mitades.

La coherencia es una medida más homogénea que el estimador sesgado clásico de autocorrelación en el sentido que todos los valores del estimador coherencia son estimados con el mismo número de muestras, en el caso de la figura anterior N/2 muestras, mientras que en el caso del estimador sesgado el número de muestras con que se calcula cada valor de la autocorrelación desciende con el índice, N muestras para el valor en el origen y N/2 muestras para el valor r(N/2).

Esta propiedad no tiene relevancia en el caso de la predicción lineal clásica sobre la señal, ya que en este caso sólo se utilizan los primeros p+1 valores de la autocorrelación y usualmente p es mucho menor que N. Sin embargo, sí puede ser interesante en el caso de aplicar predicción lineal sobre la secuencia de autocorrelación, pues en este caso se utiliza una secuencia de autocorrelación de longitud no despreciable respecto a la de la trama, la mitad en el caso de la representación SMC.

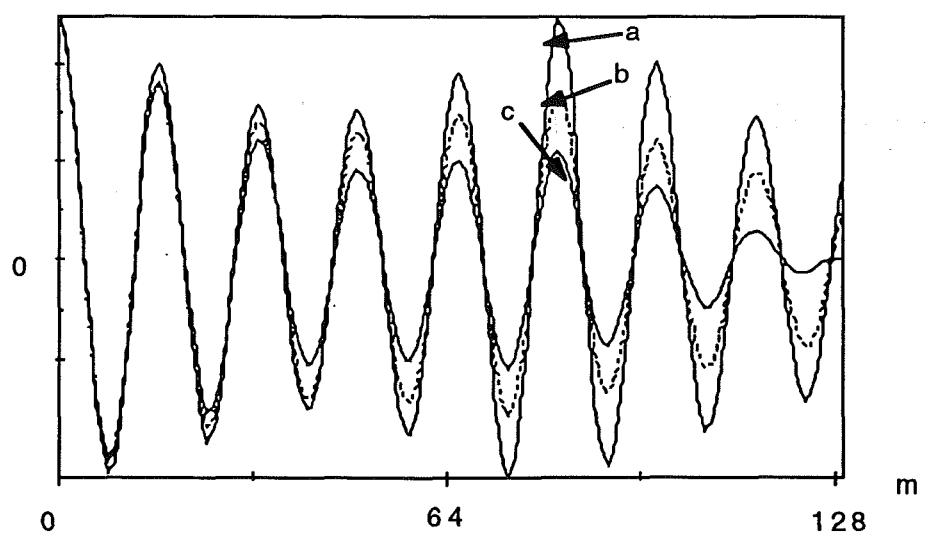


Figura 3.24. Coherencia (a) y estimador sesgado (b) calculados a partir de 254 muestras, y estimador sesgado (c) a partir de 128.

La figura 3.24 representa los primeros 128 valores de la coherencia (a) y el estimador sesgado de autocorrelación (b) estimados a partir de una trama de voz sonora de 254 muestras. Puede observarse que la coherencia marca de una forma más acentuada la periodicidad de la señal y que la diferencia entre ambos estimadores aumenta con el índice. La curva (c) corresponde al estimador sesgado para una trama de 128 muestras. En este caso, las diferencias son más notables. Debido a estas propiedades deseables de la coherencia, en las pruebas experimentales de este trabajo se ha estudiado su utilización en la técnica OSALPC.

3.6. TRANSFORMACION DE LA ESCALA DE FRECUENCIAS

A lo largo de este capítulo se ha tratado el problema de la obtención de representaciones del habla robustas al ruido mediante un análisis espectral robusto de la señal de voz desde el punto de vista del procesado de la señal. Otra aproximación al mismo problema consiste en emular las características fisiológicas o psicoacústicas del oído humano, basándose en el hecho bien conocido de que que nuestro oído parece percibir la voz mejor que cualquier máquina en presencia de ruido interferente sin un conocimiento previo de la voz ni del ruido.

Se han intentos importantes de representar el patrón de temporal de descarga de las fibras del nervio auditivo mediante complejos modelos computacionales [Gui86] [Sen88], incorporando las no-linealidades y la resolución no uniforme en frecuencia que son característicos del procesado auditivo humano. Sin embargo, tales modelos son demasiado costosos desde el punto de vista computacional y no todas las funciones que incorporan son significativas para el reconocimiento en entornos ruidosos.

Por ello, en este trabajo no se ha contemplado el diseño de modelos auditivos completos. En su lugar se han considerado evidencias auditivas que puedan incorporarse fácilmente a la etapa de parametrización del sistema de reconocimiento sin ocasionar un gasto computacional excesivo.

Algunas evidencias auditivas se han aplicado tradicionalmente en la representación de la señal de voz en reconocimiento del habla y, por tanto, estaban incorporadas en nuestro sistema de reconocimiento básico (ver capítulo 6). Así, por ejemplo, en la inmensa mayoría de los sistemas de reconocimiento se realiza un análisis espectral de la señal de voz, lo cual está en consonancia con la supuesta insensibilidad del oído a la fase de la transformada de Fourier localizada de la señal de

voz. También es bastante común la utilización como vector de parámetros acústicos de los valores iniciales del cepstrum, serie de Fourier del logaritmo del espectro, lo cual supone una compresión logarítmica en intensidad para cada frecuencia análoga a la que se produce en nuestro oído. Si se ha realizado un análisis de predicción lineal sobre la señal de voz, los coeficientes cepstrales pueden calcularse eficientemente a partir de una recursión que los relaciona los coeficientes de predicción.

Por otro lado, como se verá en el siguiente capítulo, la pronunciada sensibilidad del oído a la derivada del espectro puede modelarse en el sistema de reconocimiento mediante una medida de distorsión adecuada.

En este apartado se aborda la posibilidad de realizar una transformación de la escala de frecuencias que aproxime la sensibilidad logarítmica en frecuencia del oído. Una aproximación a la escala logarítmica de percepción del oído es la escala Mel. Tras estudiarse la aplicación directa de esta escala, se verá que puede implementarse eficientemente mediante una transformación bilineal en el plano de frecuencias complejas.

En el capítulo 6, se estudiará el comportamiento de la transformación bilineal de frecuencias en reconocimiento de habla ruidosa. Se espera que la aplicación de esta transformación de frecuencias robustezca el vector de parámetros frente al ruido aditivo de banda ancha, ya que la transformación bilineal expande la zona de bajas frecuencias, zona en que la señal de voz tiene más energía y, por tanto, es más robusta a este tipo de ruido.

Escala Mel

Una aproximación a la escala logarítmica de percepción del oído humano es la Mel, dada por la relación

$$m = 6 \log \left[(f/600) + \sqrt{1+(f/600)^2}\right], \qquad (3.94)$$

donde f está en Hz (escala lineal) y m en Barks (escala Mel). Por tanto, una distribución lineal de frecuencias en la escala Mel corresponde a una distribución logarítmica de frecuencias en la escala lineal y, por tanto, a realizar un muestreo más frecuente en la zona de bajas frecuencias que en la zona de altas frecuencias.

La obtención del espectro de una señal en la escala Mel requiere el muestreo uniforme en la escala Mel, lo cual puede realizarse a través de un banco de filtros paso-banda con anchos de banda distribuidos uniformemente en la escala Mel sobre el rango de frecuencias deseado. A partir de estas muestras del espectro se pueden obtener los coeficientes cepstrales utilizando transformada discreta de Fourier inversa.

En la figura 3.25 se muestra el proceso de obtención de los coeficientes cepstrales en la escala Mel $\hat{c}(n)$ a partir de la señal x(n) y a partir de los coeficientes cepstrales obtenidos en la escala lineal c(n).

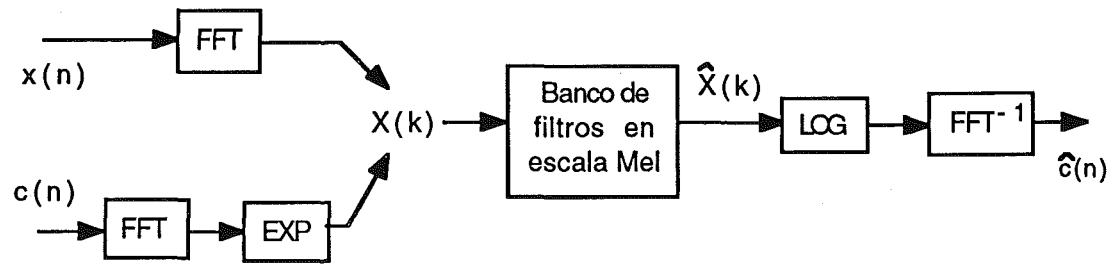


Fig 3.25 . Obtención de los coeficientes cepstrales en la escala Mel

Un método como el descrito fue utilizado por Davis [Dav80] para obtener los coeficientes cepstrales en la escala Mel a partir de la señal de voz usando un banco de filtros triangulares con anchos de banda distribuidos uniformemente en la escala Mel.

Transformación bilineal

Una alternativa al proceso descrito de transformación de los coeficientes cepstrales a la escala Mel fue propuesto por Shikano y utilizada por K.F. Lee [Lee88a], basándose en la aproximación de la escala Mel mediante una transformación bilineal.

La transformación bilineal [Opp75] es una transformación definida sobre el plano de frecuencias complejas z, que realiza una transformación no lineal del eje de frecuencias y ha sido utilizado en el diseño de filtros digitales a partir de prototipos paso-bajo.

La propiedad fundamental de la transformación bilineal es que convierte la circunferencia de radio unidad del plano z en otra circunferencia de radio unidad en un nuevo plano complejo Z de forma que la correspondencia angular entre los dos planos no es lineal. Su expresión exacta es

$$Z^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}, \qquad |\alpha| < 1 \qquad (3.95)$$

donde α es un parámetro que controla la transformación.

Evaluando (3.95) sobre la circunferencia de radio unidad del plano z se llega fácilmente a la expresión

$$\text{tg}\ \omega = \frac{(1-\alpha^2)\text{sen}\phi}{-2\alpha+(1+\alpha^2)\cos\phi}, \qquad (3.96)$$

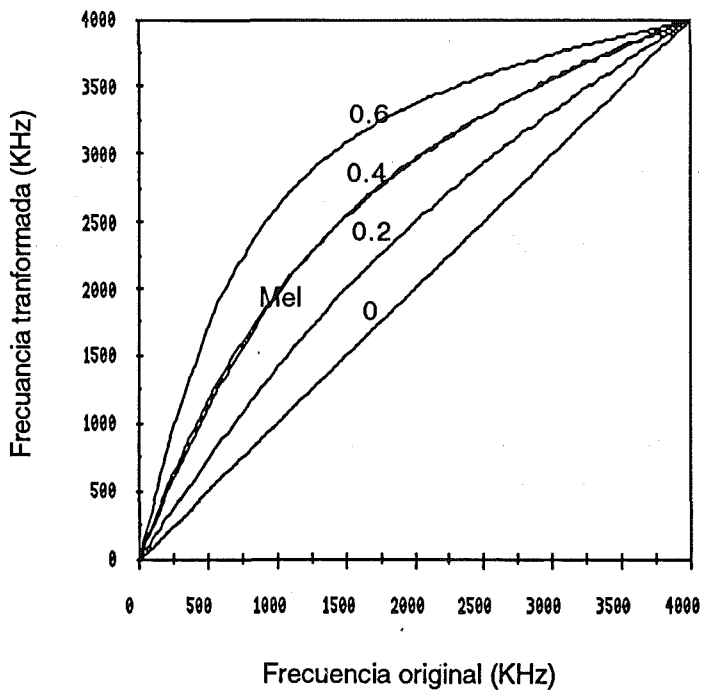


Fig. 3.26. Correspondencia entre las escalas Mel y bilineal

donde $z = e^{j\phi}$ y $Z = e^{j\omega}$, que relaciona la frecuencia original ϕ con la frecuencia transformada ω.

La compresión-expansión del eje de frecuencias realizada por la transformación bilineal está controlada por el parámetro α. Tal como se muestra en la figura 3.26, valores positivos de α producen una expansión de la zona de bajas frecuencias y una compresión de la zona de altas frecuencias. También puede apreciarse que, para una frecuencia de muestreo de 8 kHz, la transformación bilineal con $\alpha = 0.4$ corresponde aproximadamente a la escala Mel.

En la figura 3.27 se puede observar este efecto de expansión de la zona de bajas frecuencias y compresión de la zona de altas frecuencias sobre un espectro LPC de un segmento de voz sonora para distintos valores positivos del parámetro α.

La principal ventaja de la transformación bilineal respecto a la transformación directa expuesta en la figura 3.25 es que permite obtener una expresión matricial para la transformación eficiente de los coeficientes cepstrales.

Teniendo en cuenta que en reconocimiento del habla sólo se utilizan los coeficientes cepstrales de índice $n \geq 1$, los coeficientes cepstrales transformados $\hat{c}(n)$ se pueden obtener a partir de los correspondientes a la escala lineal de frecuencias c(n) aplicando la expresión

$$\hat{c}(n) = \sum_{k=1}^{\infty} c(k)\ W(n,k) \qquad n \geq 1, \qquad (3.97)$$

donde $\{W(n,k)\}_{n,k\geq 1}$ es una matriz de transformación lineal cuyos elementos pueden calcularse eficientemente mediante la siguiente recursión [Seg91]

$$W(1,k) = k\ \alpha^{k-1}\ (1-\alpha^2) \qquad k \geq 1 \qquad (3.98)$$

$$W(n,1) = (-\alpha)^{n-1}\ (1-\alpha^2) \qquad n \geq 2 \qquad (3.99)$$

$$W(n,k) = \frac{1}{(k-1)}\left[(n+k-1)\ \alpha\ W(n,k-1) + (n-1)\ W(n-1,k-1)\right]$$

$$n \geq 2,\ k \geq 2. \qquad (3.100)$$

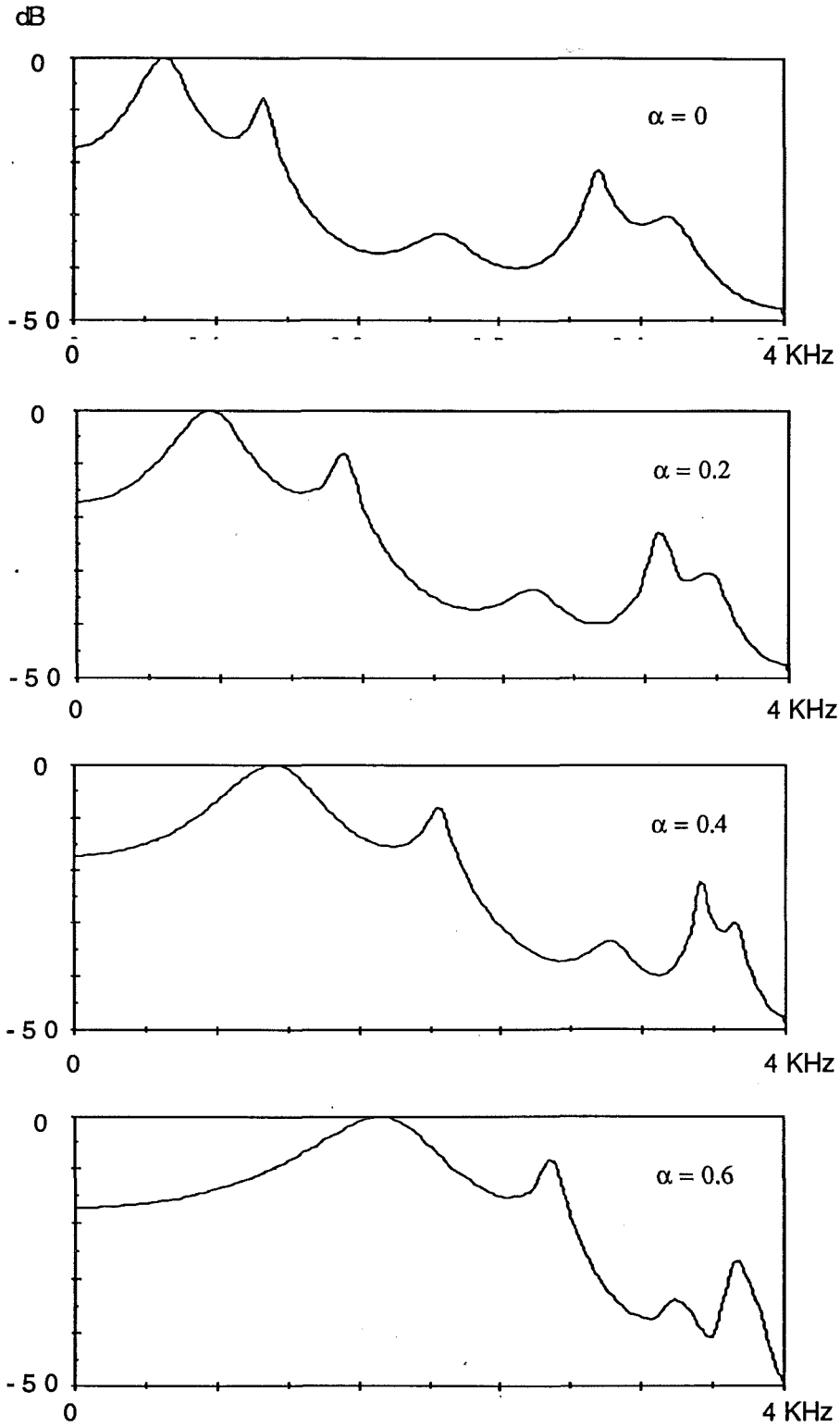


Fig. 3.27. Aplicación de la transformación bilineal a un espectro LPC correspondiente a un segmento de voz sonora