2. INTRODUCCIÓN TEÓRICA DE LOS MÉTODOS.

2.1. LA VOZ.

2.1.1. PRODUCCIÓN Y MODELADO DE LA VOZ.

Origen de las señales de voz.

La forma de onda de la voz es una onda de presión producida mediante movimientos de estructuras anatómicas que constituyen el sistema de producción de voz del ser humano.

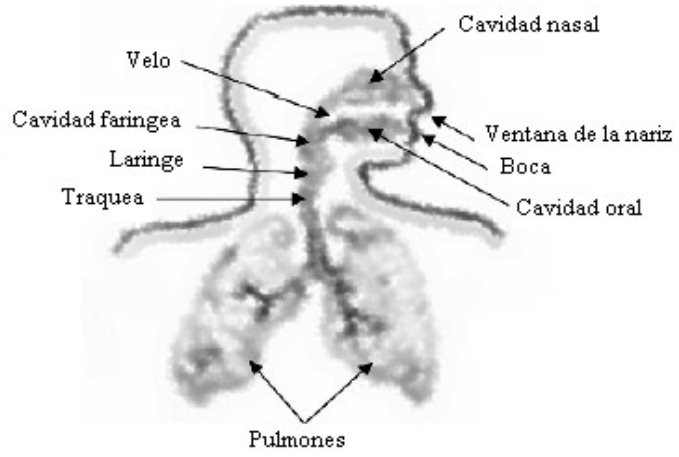

Figura 2: Sistema de producción de voz humano

Básicamente es una onda acústica emitida desde la nariz y la boca cuando el aire es expulsado por los pulmones, resultando en un flujo de aire modificado por constricciones en el interior del cuerpo humano. Este proceso puede interpretarse como una función de filtrado. El filtro estaría constituido por las cavidades nasal, oral y faríngea, y la excitación sería el aire procedente de los pulmones. Nos referimos con tracto vocal a la agrupación de la cavidad faríngea y la cavidad oral.

La forma del tracto vocal cambia continuamente, dando lugar a un filtro con respuesta variante en el tiempo. Las frecuencias de resonancia del tracto vocal es lo que se conoce como formantes, y dependen de la forma y dimensiones del tracto vocal.

En el interior de la laringe se encuentran las cuerdas vocales. Son unas bandas de músculo que se abren y se cierran durante el proceso de producción de la voz. La rapidez con la que estas se cierran y se abren es exclusiva de cada individuo.

Clasificación de las señales de voz.

Existe una distinción basta para las señales de voz, que las divide en sordas y sonoras. Los sonidos sonoros se generan cuando las cuerdas vocales se abren y se cierran, interrumpiendo de forma periódica el aire procedente de los pulmones. Con las cuerdas vocales abiertas el aire pasa a través del tracto vocal creando sonidos sordos. Los sonidos sonoros se caracterizan, por tanto, por la presencia de periodicidad en la señal, con una frecuencia fundamental llamada frecuencia de pitch. Los rangos varían notablemente entre hombres y mujeres. Los sonidos sordos no presentan ninguna periodicidad.

Es necesario indicar que esta clasificación es bastante rígida, ya que en la voz existen situaciones de transición en la que no se puede aplicar este criterio.

En general en los codificadores de voz la señal se procesa en tramas, segmentos de voz de longitud finita lo suficientemente pequeños para que las propiedades estadísticas se mantengan constantes en el intervalo. Esta longitud ronda entre los 20 y 30 milisegundos.

Modelado del proceso de producción de voz.

El sistema de producción de voz del ser humano se puede modelar de la siguiente manera: podemos considerar los pulmones como la fuente de excitación (ruido blanco), y el tracto vocal se puede asociar a un filtro variante en el tiempo. Estos parámetros pueden estimarse mediante una técnica de identificación de sistemas llamada predicción linear. La asunción del modelo es que la

distribución de la energía en frecuencia se debe exclusivamente al filtro, mientras que los pulmones producen un ruido blanco con un espectro plano. El sistema auditivo humano no es tan sensible a la fase como a la magnitud del espectro.

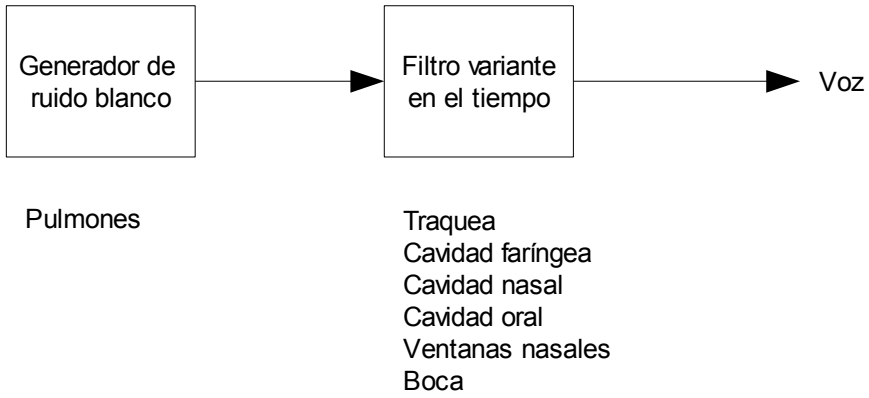

Figura 3: Modelado del sistema de producción de voz humano

Estructura general de un codificador de voz.

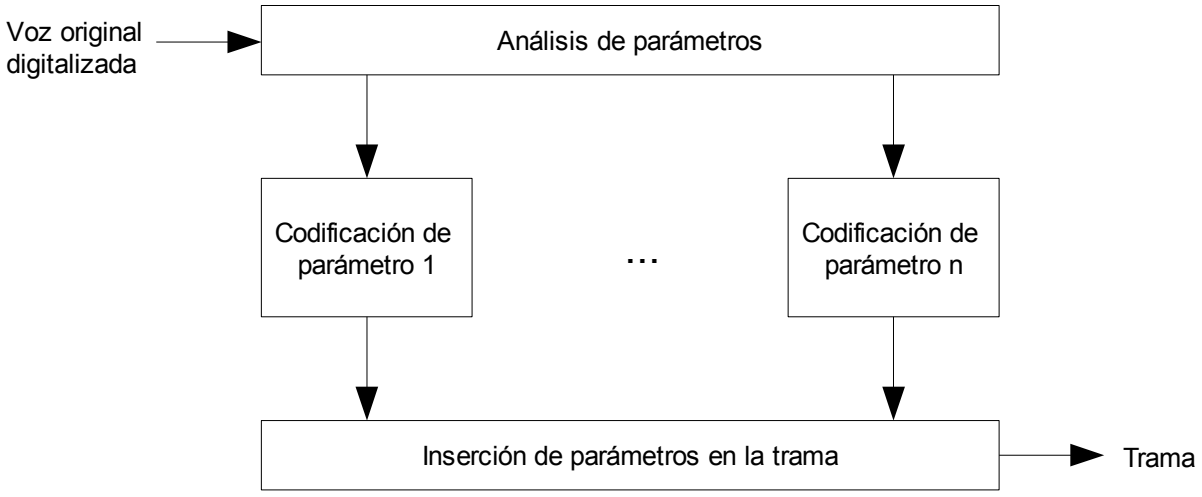

Figura 4: Estructura genérica de un codificador de voz I

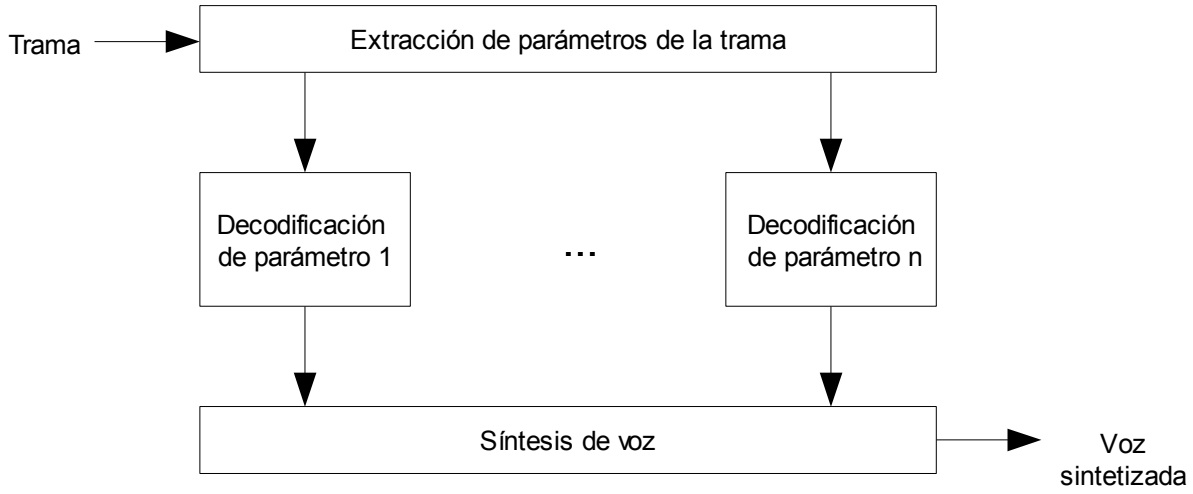

Figura 5: Estructura genérica de un codificador de voz II

El codificador partirá de una versión digitalizada de la fuente de voz. Según el tipo de codificador, se llevará a cabo un proceso de análisis para la extracción de los parámetros necesarios para la reconstrucción en el otro extremo. Una vez seleccionados, serán codificados e insertados en una trama o flujo de bits, que servirá de fuente para el esquema del decodificador.

El decodificador separará los parámetros de la trama y los decodificará. A partir de ellos sintetizará la señal de voz.

2.1.2. CLASIFICACIÓN DE LOS CODIFICADORES DE VOZ.

Antes de comenzar con el desarrollo de los métodos comentaremos brevemente algunas clasificaciones de codificadores de fuente, y veremos dónde se sitúan nuestros esquemas de codificación. Existen varias:

1. Clasificación por técnica de codificación.

   - Codificadores de forma de onda. Su propósito es el de preservar la "forma" de la señal.

Los esquemas basados en DCT y ADPCM pertenecen a este grupo.

- Codificadores paramétricos. Están basados en un modelo, producto de un estudio de las características de la fuente. No persiguen representar de manera fidedigna a la señal, sino sus características. El esquema basado en LPC pertenece a este grupo.

- Codificadores híbridos. Como su nombre indica, combina características de los codificadores anteriores: por una parte trabaja sobre un modelo, por otra persigue imitar la señal. El esquema basado en CELP pertenece a este grupo.

2. Clasificación modo simple y multimodo.

   - Modo simple. Estos codificadores trabajan a una tasa fija para un flujo de bits dado. A este grupo pertenecen todos los esquemas que son objeto de este proyecto.

   - Multimodo. Los codificadores multimodo se caracterizan por poder variar su tasa de bit, dependiendo de las características estadísticas de la voz o del estado de la red en ese momento. De esta manera pueden adaptarse a unas condiciones en concreto. Si por ejemplo, la red estuviera muy cargada, podrían comportarse como un codificadores de tasa baja, y viceversa.

3. Clasificación por tasa.

   - Tasa alta. Son codificadores cuya tasa supera los 15 kbps. El esquema basado en ADPCM y el Codificador Armónico pertenecen a este grupo.

   - Tasa media. Su tasa se halla comprendida entre los 5 y 15 kbps. El esquema basado en CELP pertenece a este grupo.

   - Tasa baja. Su tasa queda por debajo de los 5 kbps. El esquema basado en LPC pertenece a este grupo.

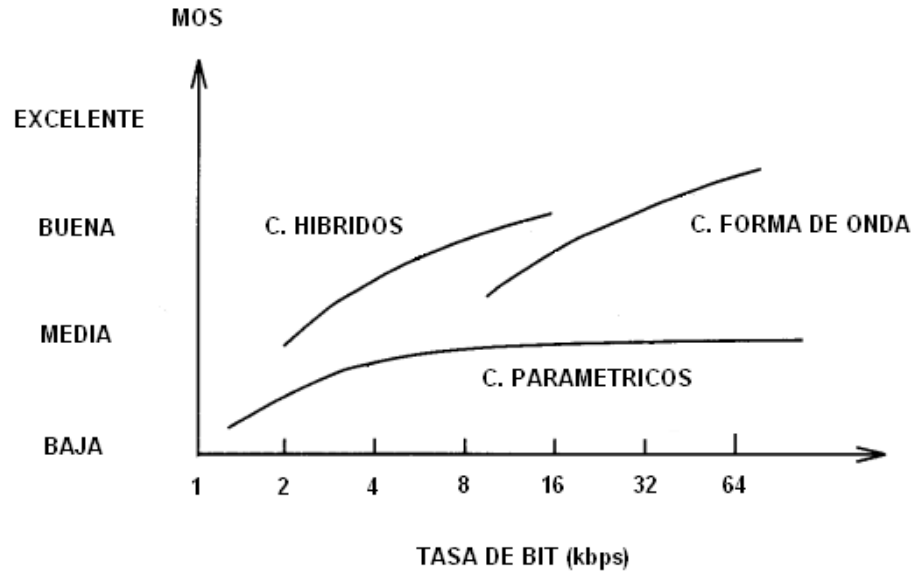

Figura 6: Codificadores: Tasa y calidad

Una vez vistas a grosso modo las características de los esquemas pasaremos a verlos un poco más en profundidad, unos más que otros dada su complejidad.

2.2. CODIFICADOR ARMÓNICO.

Es un codificador basado en la aplicación directa de la transformada discreta de Fourier (a partir de este momento DFT).

Este método esta basado en la propiedad cuasiestacionaria de la voz, asumiendo que pequeños tramos de voz en el tiempo pueden ser representados mediante la DFT. Durante el proceso de análisis para cada tramo enventanado se eligen un conjunto de sinusoides para representar la voz. Los parámetros de estas sinusoides se seleccionan sobre el espectro: las tripletas {Amplitud, Frecuencia, Fase}. La compresión se traduce en limitar el numero de tripletas que enviaremos al decodificador.

Un criterio muy básico para ello consiste en tomar los N puntos con mayor valor, limitándonos así a los máximos de la magnitud del espectro.

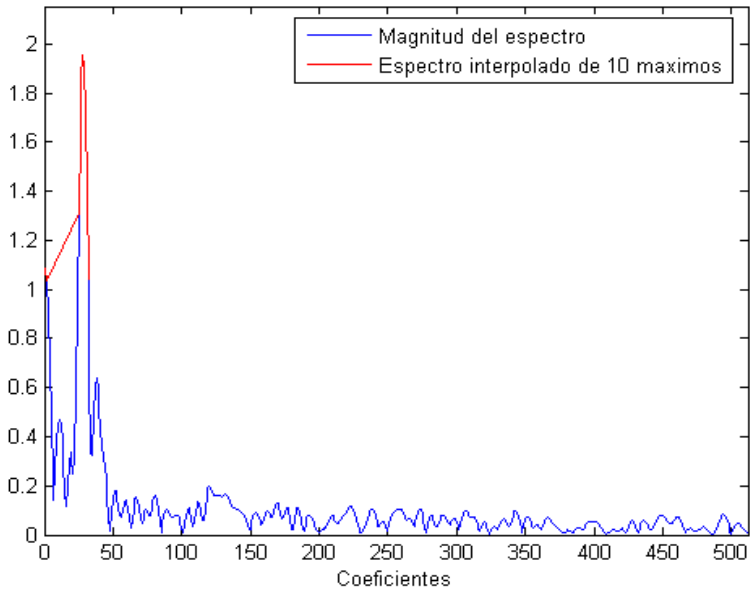

Figura 7: Selección de diez máximos sobre el espectro de la DFT

La ventaja de este criterio es que la SNR del tramo reconstruido es la mayor respecto a cualquier otro método. La desventaja es que, de esta manera es probable que estemos perdiendo la "forma" de la magnitud del espectro (los formantes).

El proceso de síntesis para cada trama se reduce a sumar las sinusoides generadas a partir de las tripletas procedentes del codificador.

En sí mismo, el codificador armónico no puede clasificarse a priori como de forma de onda o paramétrico, sino que dependerá del criterio de selección de las tripletas. El ejemplo anterior correspondería a un codificador de forma de onda ya que intenta imitar (con un número limitado de puntos) la forma de la señal.

2.3. CODIFICADOR BASADO EN LA DCT.

Es un codificador sencillo que usa la transformada discreta del coseno (a partir de este momento DCT).

La DCT es una transformada de frecuencia basada, a su vez, en la DFT, mejorando su compactación de energía en el caso de señales altamente correladas (por ejemplo, imágenes), sin sacrificar la eficiencia computacional. Esto es posible debido a que la DFT se calcula partiendo de una señal periódica con discontinuidades artificiales. Estas discontinuidades aumentan la energía en alta frecuencia, reduciendo la compactación. La DCT salva este problema, ya que se calcula sobre una señal sin discontinuidades, mejorando la compactación.

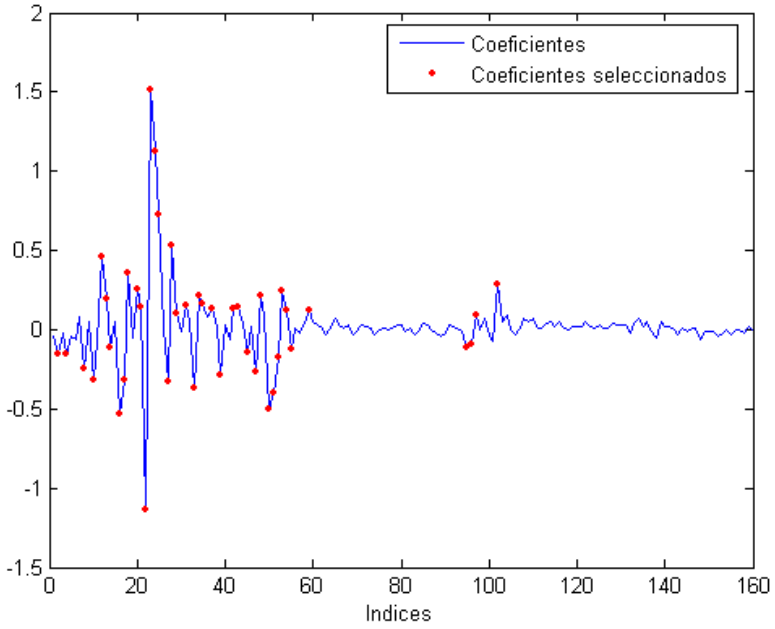

Figura 8: Selección de coeficientes de la DCT

Las DCT compacta la señales de voz en un conjunto reducido de índices. En el proceso de análisis se han seleccionado aquellos que superen un umbral fijo de magnitud para todas las tramas.

El proceso de síntesis consiste en reconstruir los vectores de los coeficientes DCT, ubicando en sus respectivos índices los coeficientes seleccionados procedentes del codificador, y al resto asignándoles el valor cero. Sólo queda aplicar la transformada inversa del coseno (a partir de ahora IDCT) a estos vectores, obteniéndose así la voz sintetizada.

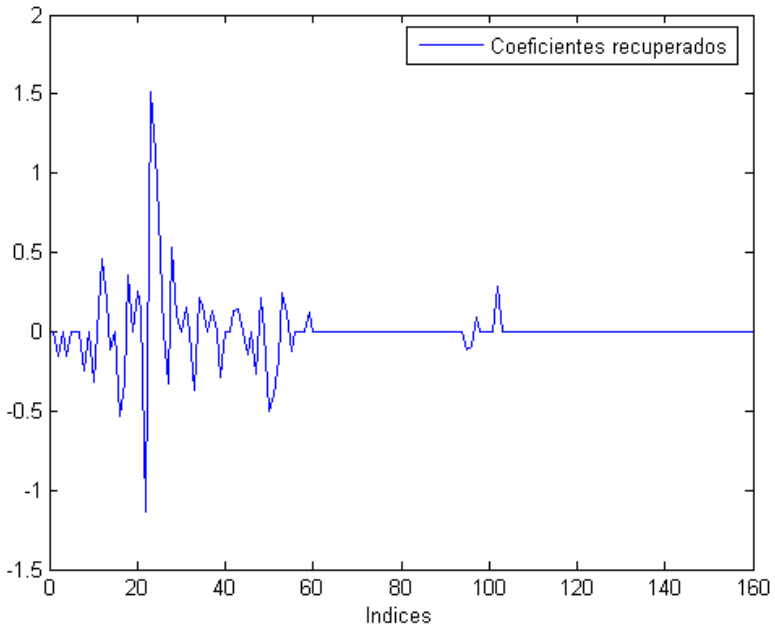

Figura 9: Reconstrucción de la DCT

2.4. CODIFICADOR ADPCM.

2.4.1. PCM (Modulación por codificación de pulsos).

Podría entenderse como un proceso de cuantización de muestras en el tiempo discreto, de tal manera que ambas magnitudes, tiempo y amplitud, son representadas con precisión finita.

A continuación presentamos un ejemplo de cuantización de una señal con un tiempo de muestreo de 0.1 segundos y 3 posibles niveles de amplitud:

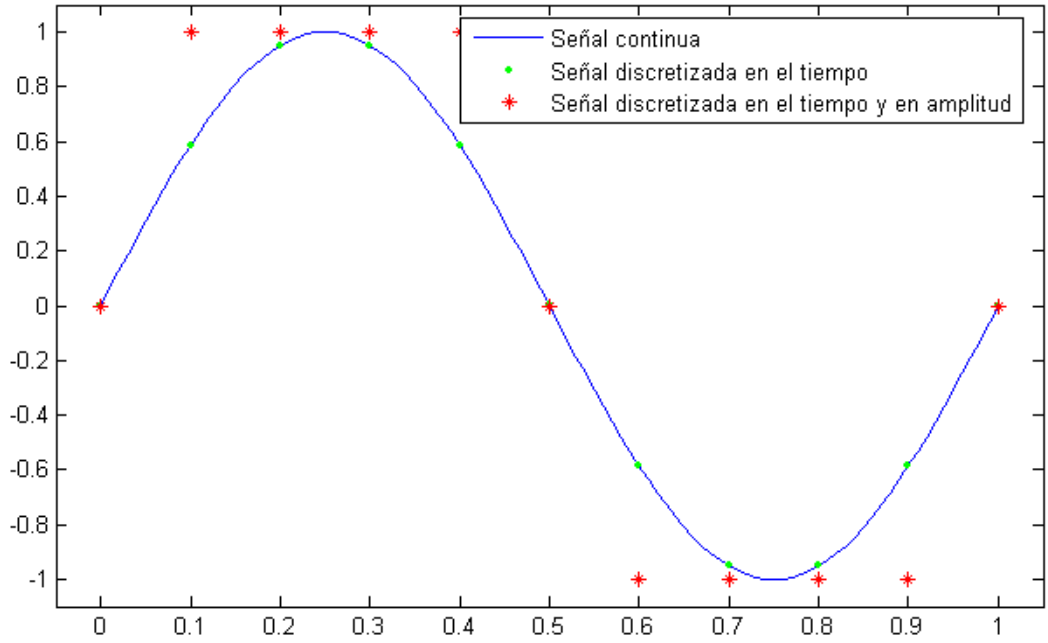

Figura 10: Ejemplo de discretización en tiempo y amplitud

Es una codificación de propósito general, válida para cualquier tipo de señal y por tanto, ineficiente, pero nos sirve como un referente para comenzar y comparar.

La cuantización de muestras consiste en aproximar todos los valores continuos contenidos en un intervalo a un valor digital que los representa a todos. Se define el paso de cuantización como el salto analógico entre dos valores digitales consecutivos.

Existen dos tipos, dependiendo si el paso de cuantización permanece constante o no:

1. Cuantización uniforme. Presenta una función de cuantización 'lineal':

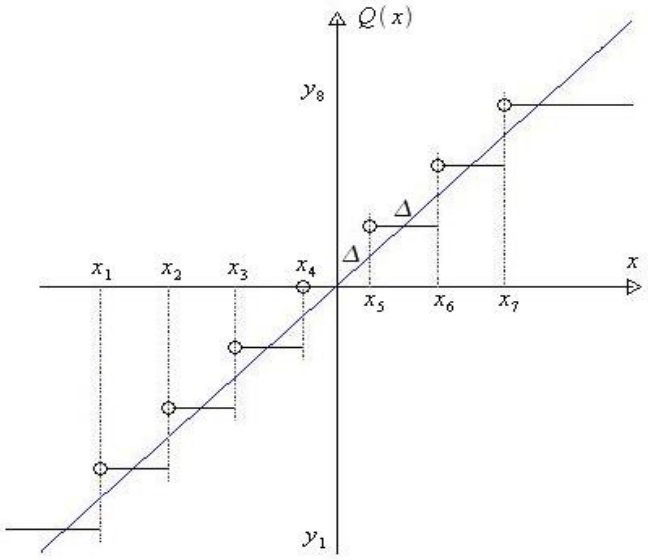

Figura 11: Función de transferencia de un cuantizador uniforme

2. Cuantización no uniforme. En este caso la variable a cuantizar se transforma mediante el uso de funciones no lineales, después se cuantiza con un cuantizador uniforme y al resultado se le aplica la inversa de la función lineal aplicada en el primer paso. De esta manera se consigue proporcionar más precisión a cierto rango de amplitudes de la señal original. Existen dos funciones logarítmicas que han sido muy utilizadas en telefonía digital:

   1) Ley μ. Se caracteriza por la siguiente función:

$$Q(x) = A\frac{\ln(1+\mu|x|/A)}{\ln(1+\mu)}sgn(x) \quad para\ |x| \le A$$

   2) Ley A. Se caracteriza por la siguiente función:

$$Q(x) = \begin{cases} \dfrac{A_o|x|}{1+\ln A_o}sgn(x) & para\ |x| \le \dfrac{A}{A_o} \\ \dfrac{A(1+\ln(A_o|x|/A))}{1+\ln A_o}sgn(x) & para\ \dfrac{A}{A_o} \le |x| \le A \end{cases}$$

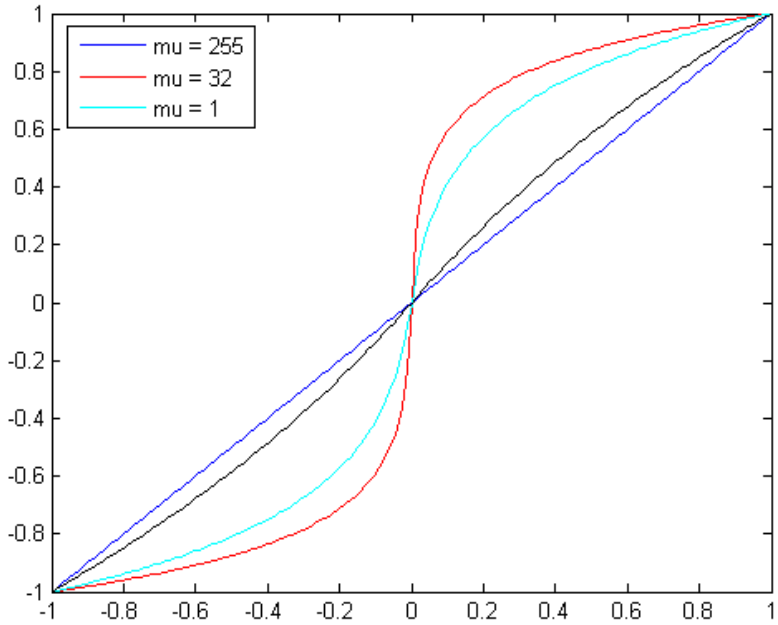

Figura 12: Característica de la ley μ

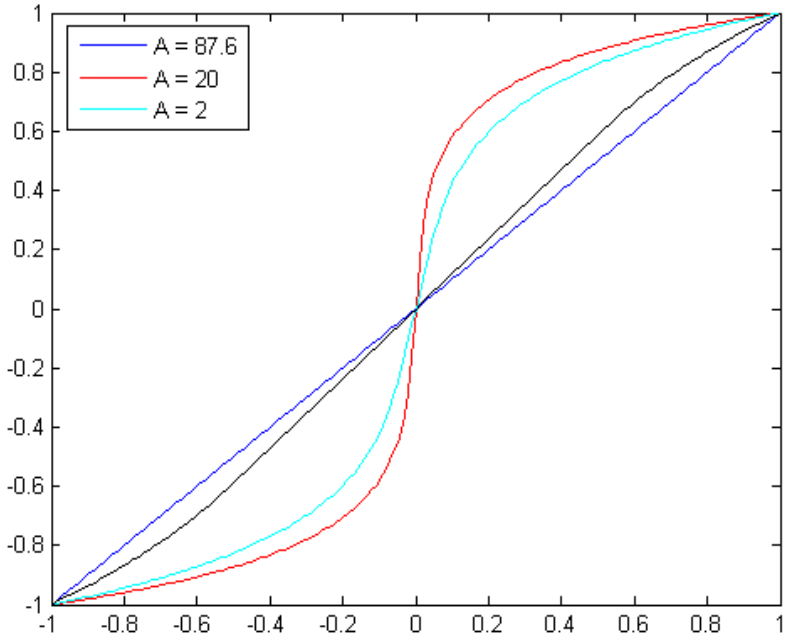

Figura 13: Característica de la ley A

2.4.2. DPCM (Modulación por codificación de pulsos diferencial).

DPCM está basada en la cuantización no de la señal en sí misma, sino en la de un error de predicción. Este método se usa en escenarios donde la señal a cuantizar presenta cierto grado de correlación entre sus muestras vecinas o correlativas. Si podemos predecir la muestra actual a partir de las pasadas, podemos obtener la señal de error, que se caracteriza por un rango dinámico menor.

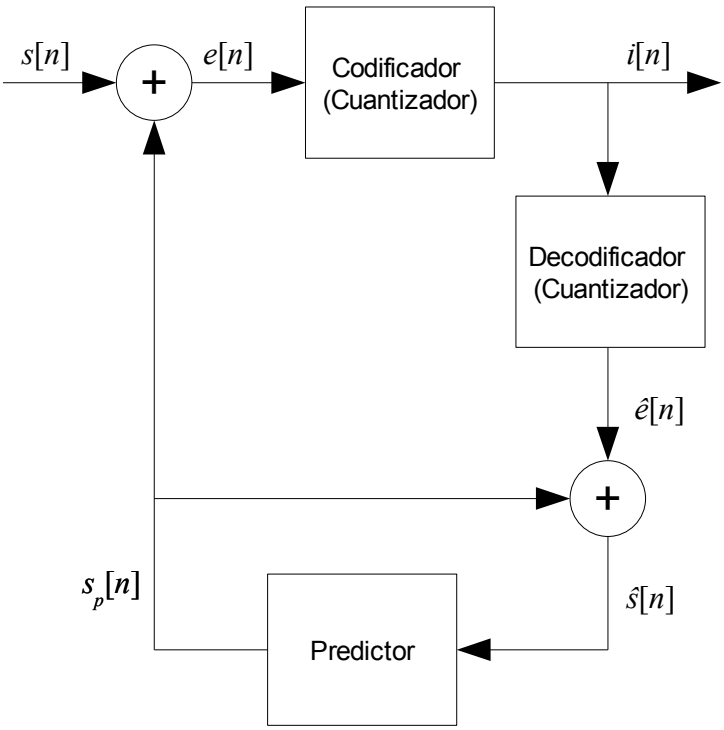

Figura 14: Codificador DPCM AR

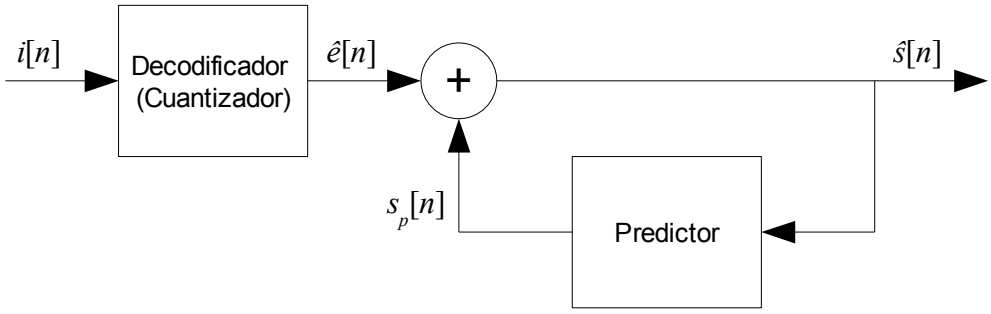

Figura 15: Decodificador DPCM AR

En este codificador la predicción esta basada en un conjunto de muestras cuantizadas. Podrían conseguirse mejores resultados usando las muestras originales no cuantizadas. Se usa la aproximación debido a que el codificador no tiene acceso a la entrada original y ambos, codificador y decodificador, deben mantenerse sincronizados

El esquema anterior esta orientado a un modelo autoregresivo (AR), donde se utiliza las muestras cuantizadas de la señal de entrada. Otra posibilidad sería usar un modelo MA, donde la entrada al predictor fuera la señal de error. Su rendimiento es menor que el basado en el modelo AR, pero también es más robusto frente a los errores de canal, debido a que el modelo AR perpetuaría esos errores, mientras que en el modelo MA solo afectaría a un número de muestras finitas.

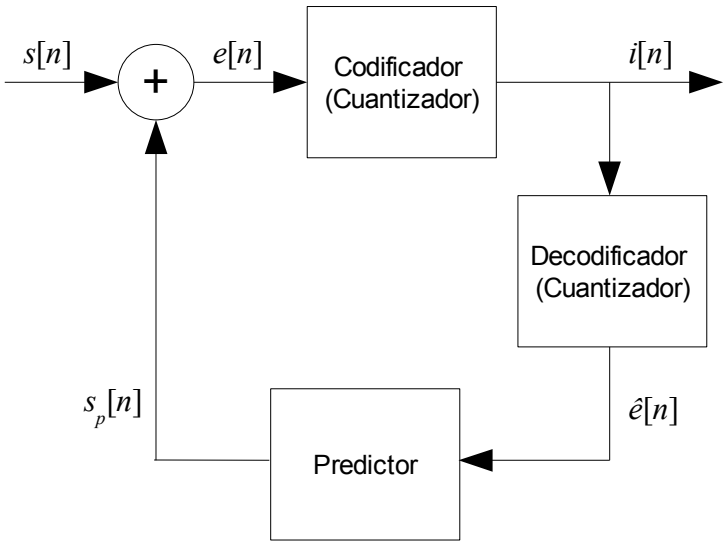

Figura 16: Codificador DPCM MA

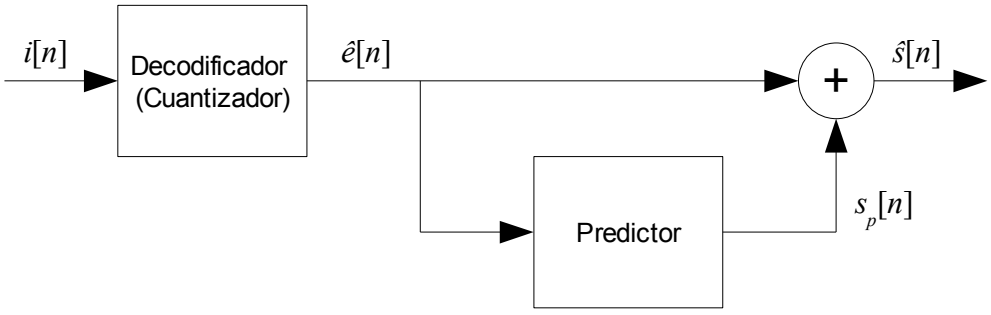

Figura 17: Decodificador DPCM MA

2.4.3. APCM (Modulación por codificación de pulsos adaptativa).

APCM supone la inclusión de sistemas que se adapten a la señal de entrada. Estos sistemas son fundamentales cuando tratamos con señales no estacionarias, como la voz. Existen dos esquemas básicos que comentaremos a continuación:

1. Cuantizador de ganancia adaptativa hacia adelante. Este esquema permite controlar de forma precisa en nivel de ganancia, pero como contrapartida tiene un exceso envío de información respecto al esquema hacia atrás.

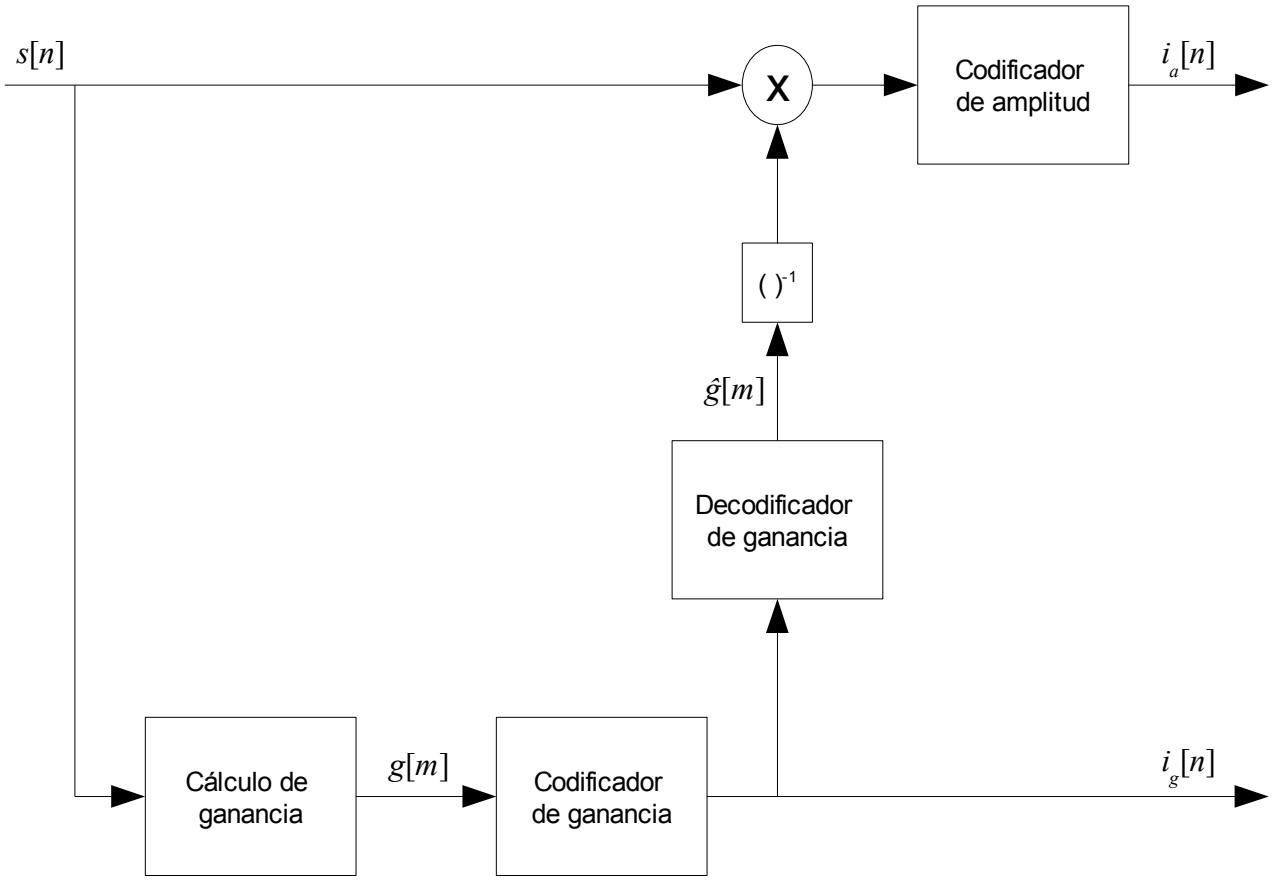

Figura 18: Codificador APCM hacia adelante

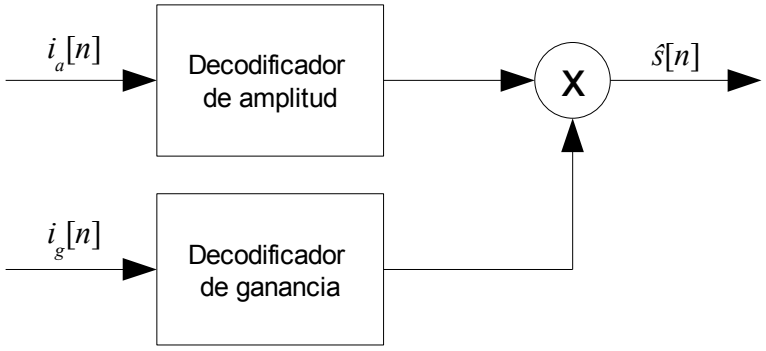

Figura 19: Decodificador APCM hacia adelante

La señal de entrada se divide en tramas. Para cada una, se estima una ganancia. El siguiente paso es reescalar las muestras de la trama dividiendo por la ganancia. De esta manera estamos reduciendo el rango dinámico de la señal de entrada. Si se produce algún tipo de error en la transmisión en un instante dado éste sólo afectará al grupo de tramas implicadas, no se transmitirá de manera indefinida a las siguientes.

La ganancia puede calcularse de forma sencilla, por ejemplo:

$$g[n]=k_1 max(x|n|)+k_2$$

2. Cuantizador de ganancia adaptativa hacia atrás. Igual antes, se trata de calcular una ganancia con el objetivo de normalizar las muestras de la señal de entrada y disminuir su margen dinámico. En este caso la ganancia se estima a la salida del codificador. Como ventaja no tiene que enviar el valor de la ganancia, pero en caso de que se produzca un error en la transmisión éste afectará a la memoria del estimador de ganancia, llevando a una propagación del error hacia adelante.

   La ganancia en este caso se calcula como un valor proporcional a la estimación recursiva de la varianza de las muestras normalizadas, que se obtiene de la siguiente manera:

$$\sigma^2[n]=\alpha\sigma[n-1]+(1-\alpha)y[n]$$

con $0 < \alpha < 1$. La ganancia queda por tanto:

$$g[n] = k_1 \sigma^2[n] + k_2$$

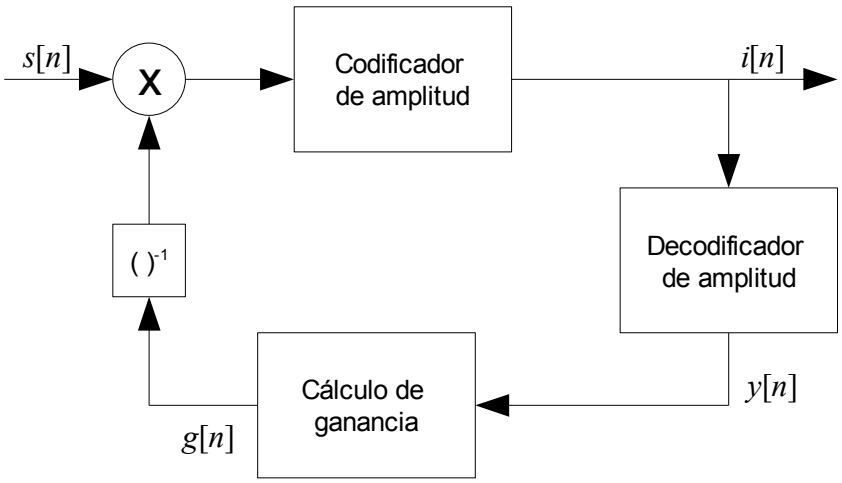

Figura 20: Codificador APCM hacia atrás

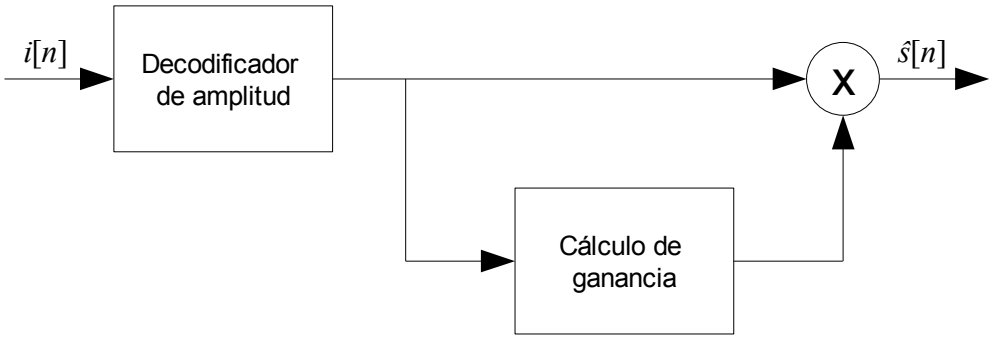

Figura 21: Decodificador APCM hacia atrás

2.4.4. ADPCM (Modulación por codificación de pulsos diferencial adaptativa).

DPCM esta basada en un predictor y en un cuantizador fijos. Estos sistemas se pueden mejorar introduciendo adaptación en uno de los dos módulos, o en ambos. Es lo que se conoce

como ADPCM. A continuación comentaremos dos esquemas con predictores y cuantizadores adaptativos, uno con adaptación hacia adelante y otro con adaptación hacia atrás:

1. ADPCM con adaptación hacia adelante. Al igual que en el esquema APCM con adaptación hacia adelante, se transmite exceso de información, en este caso no sólo la ganancia sino información del predictor. En caso de tener un predictor lineal podemos calcular sus coeficientes LPC. Por otra parte también comparte similitudes con DPCM, donde se calcula el error de predicción como diferencia entre la muestra original y la predicha. En este caso, para el cálculo de la ganancia se utiliza la trama de error de predicción en vez de la trama de la propia señal y, una vez calculada, se usa para normalizar las muestras de error. El error de predicción se usa para predecir la señal, como ocurre en los esquemas DPCM.

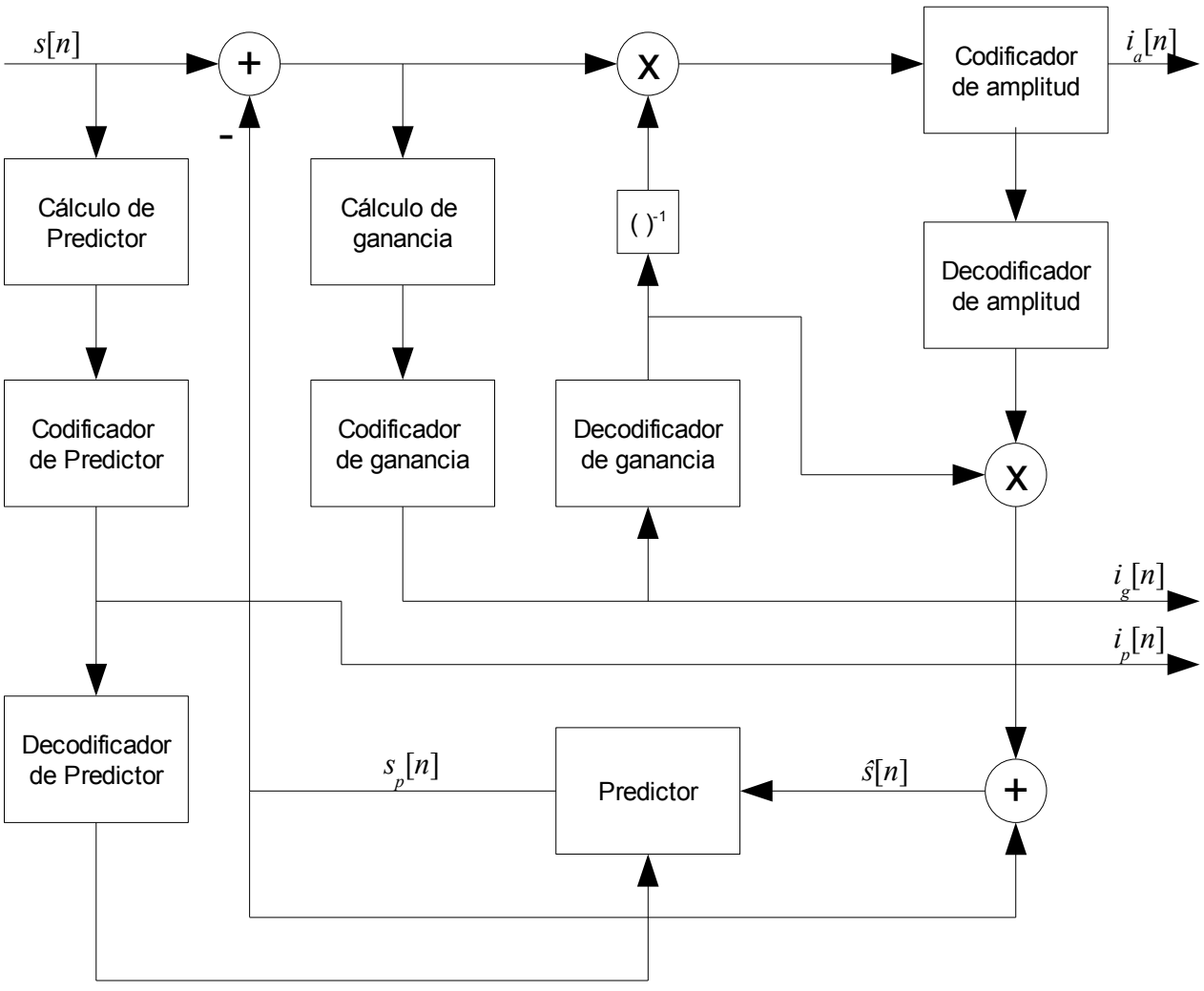

Figura 22: Codificador ADPCM hacia adelante

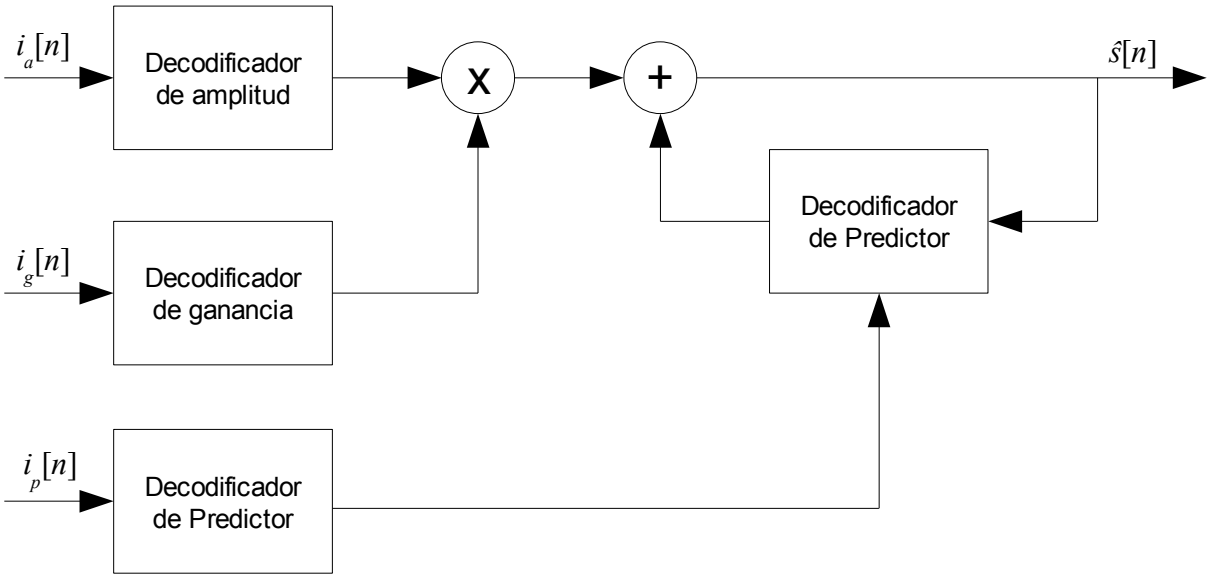

Figura 23: Decodificador ADPCM hacia adelante

2. ADPCM con adaptación hacia atrás. Al igual que en DPCM, se obtiene el error de predicción como diferencia de la muestra de la señal de entrada y la predicción, que es escalado, cuantizado y transmitido. El error normalizado y cuantizado se utiliza para el cálculo de la ganancia. Esta ganancia se usa tanto para normalizar la siguiente trama de error como para desnormalizar el error de predicción, que se usará para calcular la predicción para la siguiente trama. Para un posible predictor lineal podrían calcularse sus coeficientes LPC.

   Respecto al anterior, este esquema es más rápido (tiene menos carga computacional), envía menos datos pero es más sensible a los errores de transmisión, ya que un error puntual afectaría a la muestra actual y a las futuras.

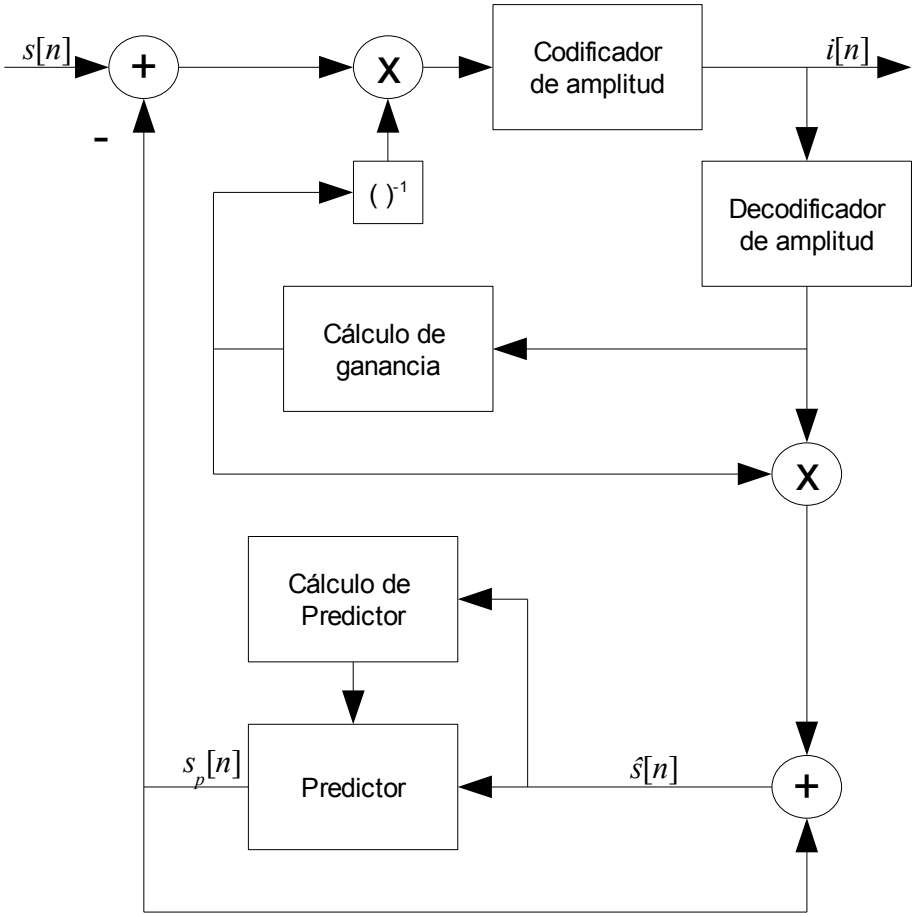

Figura 24: Codificador ADPCM hacia atrás

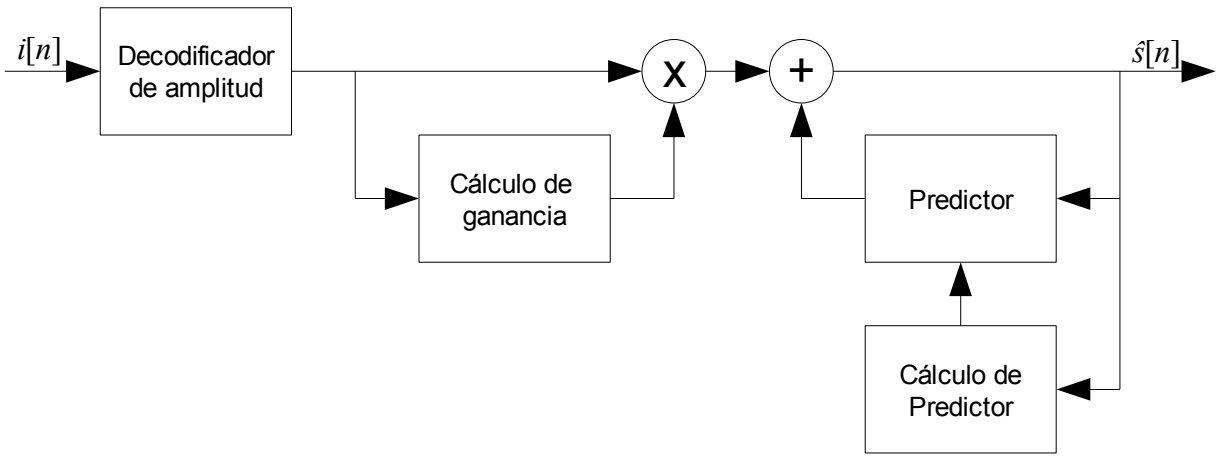

Figura 25: Decodificador ADPCM hacia atrás

2.5. CODIFICADOR LPC.

2.5.1. ANÁLISIS

El modelo físico de la producción de voz puede representarse mediante un esquema matemático:

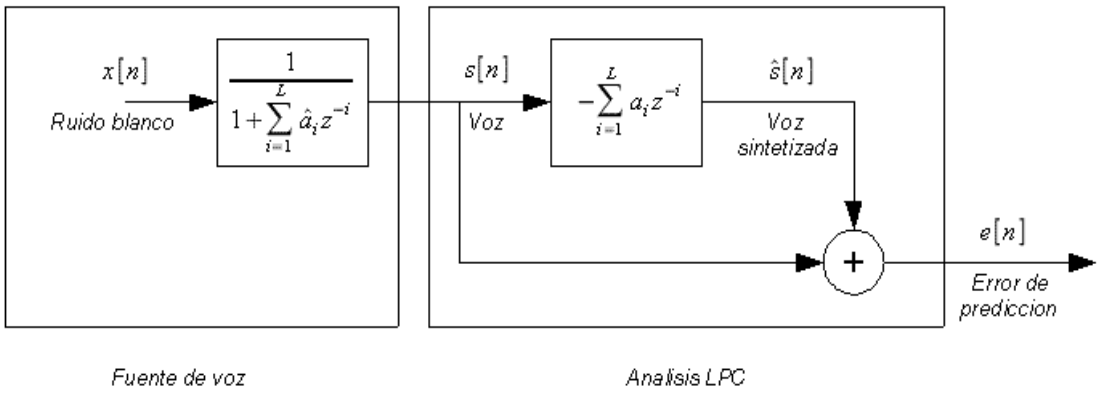

Figura 26: Predicción lineal como identificación de sistemas

En el modelo, los pulmones generan una excitación de tipo ruido blanco, y la función del tracto vocal puede asociarse a un sistema variante en el tiempo, que se aproximará por un sistema autoregresivo. Los parámetros de este sistema podremos estimarlos mediante la técnica de predicción lineal (LP).

Predicción lineal.

El principio básico de la predicción lineal consiste en que una muestra de voz puede aproximarse como una combinación lineal de muestras pasadas. Para cada trama los pesos de la combinación lineal pueden estimarse denominándose coeficientes de predicción lineal (LPC).

$$\hat{s}[n] = -\sum_{i=1}^{L} a_i s[n-i]$$

En general puede representarse una trama como la señal aproximada más un error:

$$s[n]=\hat{s}[n]+e[n]=-\sum_{i=1}^{L} a_i s[n-i]+e[n]$$

De lo anterior se puede concluir que el tracto vocal tiene la siguiente forma:

$$H(z)=\frac{1}{1+\sum_{i=1}^{L} a_i z^{-i}}$$

donde L se conoce como orden del predictor.

El siguiente paso para estimar los coeficientes consiste en minimizar el error, haciendo de esta manera que la señal original se parezca lo máximo posible a la señal por la cual queremos aproximarla.

A continuación introduciremos los cálculos matemáticos que nos llevaran a la obtención de los coeficientes:

1. Definimos el error cuadrático medio.

$$J=E\{e^2[n]\}=E\{(s[n]+\sum_{i=1}^{L} a_i s[n-i])^2\}$$

2. Minimizamos el error para cada coeficiente.

$$\frac{\partial J}{\partial a_k}=2E\{(s[n]+\sum_{i=1}^{L} a_i s[n-i])s[n-k]\}=0$$

3. Esta ecuación puede reescribirse de otra manera.

1. $$E\{s[n]s[n-k]\}+\sum_{i=1}^{L} a_i E\{s[n-i]s[n-k]\}=0$$
$$para\ k=1,2,\ldots,L$$

2. $$\sum_{i=1}^{L} a_i R_s[i-k]=-R_s[k]$$
$$para\ k=1,2,\ldots,L$$

4. O expresarse de forma matricial.

$$R_s a = -r_s$$

donde

$$R_s = \begin{pmatrix} R_s[0] & R_s[1] & ... & R_s[L-1] \\ R_s[1] & R_s[0] & ... & R_s[L-2] \\ \vdots & \vdots & \vdots & \vdots \\ R_s[L-1] & R_s[L-2] & ... & R_s[0] \end{pmatrix}$$

$$a = \begin{bmatrix} a_1 & a_2 & ... & a_L \end{bmatrix}^T$$

$$r_s = \begin{bmatrix} R_s[1] & R_s[2] & ... & R_s[L] \end{bmatrix}^T$$

5. Suponiendo que la matriz de correlación R_s tiene inversa, podemos hallar la matriz de coeficientes óptima.

$$a = -R_s^{-1} r_s$$

La solución a la ecuación anterior puede resolverse hallando la inversa de la matriz R_s. Habitualmente el cálculo de matrices inversas es un proceso costoso desde el punto de vista computacional. Existen diversos algoritmos que nos permitirán hallar el vector de coeficientes LPC aprovechando la estructura de la matriz de correlación.

Comentaremos brevemente el algoritmo de Levinson-Durbin. Partiremos del siguiente sistema:

$$\begin{pmatrix} R[0] & R[1] & \cdots & R[L] \\ R[1] & R[0] & \cdots & R[L-1] \\ \vdots & \vdots & \ddots & \vdots \\ R[L] & R[L-1] & \cdots & R[0] \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_L \end{pmatrix} = \begin{pmatrix} J \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

En la práctica los valores de la correlación se estiman a partir de las muestras de la señal y J, el error cuadrático medio es desconocido, aunque el algoritmo nos proporcionará su valor.

Levinson-Durbin permite encontrar la solución de un predictor de orden M a partir de otro de orden M-1. Se trata de un proceso donde el primer paso es la resolución del predictor de orden 0.

A partir de él, hallaremos el de orden 1, y así sucesivamente hasta alcanzar el orden deseado. El proceso se basa en dos claves:

1. La matriz de correlación contiene como subbloques todas las matrices de correlación de orden menor.

2. La matriz de correlación es de tipo Toeplitz, es decir, todos los elementos de su diagonal principal son iguales, así como los elementos de cada diagonal paralela a la principal. Si tenemos la siguiente estructura matricial:

$$\begin{pmatrix} R[0] & R[1] & \cdots & R[L] \\ R[1] & R[0] & \cdots & R[L-1] \\ \vdots & \vdots & \ddots & \vdots \\ R[L] & R[L-1] & \cdots & R[0] \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_L \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_L \end{pmatrix}$$

Entonces:

$$\begin{pmatrix} R[0] & R[1] & \cdots & R[L] \\ R[1] & R[0] & \cdots & R[L-1] \\ \vdots & \vdots & \ddots & \vdots \\ R[L] & R[L-1] & \cdots & R[0] \end{pmatrix} \begin{pmatrix} a_L \\ a_{L-1} \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} b_L \\ b_{L-1} \\ \vdots \\ b_0 \end{pmatrix}$$

Como se observa, la matriz de correlación es invariante al cambio de sus columnas y después de sus filas.

El algoritmo de Levinson-Durbin consta de un proceso iterativo-recursivo de tres pasos:

1. *Inicialización: d=0; esto nos lleva a*

$$J_0 = R[0]$$

2. *Recursión: para d=1,2,...,L*

1. Cálculo del coeficiente de reflexión.

$$k_d = \frac{1}{J_{d-1}} \left(R[d] + \sum_{i=1}^{d-1} a_i^{d-1} R[d-i] \right)$$

2. *Cálculo de los coeficientes del predictor de orden d*

$$a_d^d = -k_d$$

$$a_i^d = a_i^{d-1} - k_d a_{d-1}^{d-1} \; para \; i = 1,2,\ldots,d-1$$

hasta que d=L.

3. *Calcular el error cuadrático medio asociado al predictor de orden d.*

$$J_d = -J_{d-1}(1-k_d^2)$$

Incrementar d (d=d+1) y volver al paso 1.

De esta manera hemos obtenido los coeficientes LPC para cada trama.

Cálculo de la ganancia.

Para el cálculo de la ganancia, supondremos que la excitación del sistema *x[n]* es un ruido blanco de ganancia unidad. Por tanto:

$$Var(x[n]) = E\{x^2[n]\} = 1$$

La excitación se multiplica por una ganancia *g*. La varianza de la señal *gx[n]* es:

$$Var(gx[n]) = E\{(gx[n])^2\} = g^2$$

Por otra parte, sabemos que la varianza de la entrada al filtro LPC *gx[n]* coincide con el error cuadrático medio, tal como fue definido:

$$J_{min} = E\{(gx[n])^2\} = g^2$$

Despejando, tenemos la ganancia:

$$g = \sqrt{J_{min}}$$

El valor de J_{min} podemos obtenerlo también mediante el algoritmo de Levinson-Durbin. De esta manera hemos obtenido la ganancia de predicción para cada trama.

Predicción corta y predicción larga.

Experimentalmente puede comprobarse que el orden del predictor debe ser lo suficientemente alto para poder modelar de forma adecuada el segmento de voz considerado. El hecho de incrementar el orden de predicción nos lleva a aumentar la tasa de bit para representar los coeficientes LPC, y aumenta el coste computacional durante el análisis. Ahora bien, a partir de cierto orden, los coeficientes LPC dejan de contribuir de forma significativa al modelo. Estos coeficientes pueden ser eliminados, llevándonos a un esquema más eficiente. Esta es la idea de la predicción linear larga.

El análisis quedaría de la siguiente manera:

1. En primer lugar nos encontramos con el predictor corto, que es un predictor como el que hemos estudiado anteriormente con un orden entre 8 y 12, cuya misión sera eliminar la redundancia entre muestras cercanas.

2. El siguiente módulo sería el predictor largo, que actuaría sobre muestras separadas entre sí por un periodo de pitch.

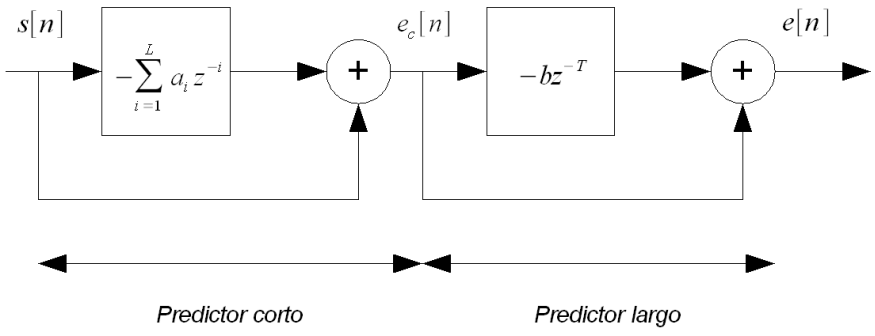

Figura 27: Conexión de un predictor corto y un predictor largo

El predictor largo tendría la siguiente función de transferencia:

$$H(z) = 1 + bz^{-T}$$

Los dos parámetros sin especificar son la ganancia de predicción larga *b* y el periodo de pitch *T*.

Análisis de predicción larga.

La predicción larga consiste en predecir la muestra actual a partir de una muestra pasada en el tiempo a la distancia de un periodo de pitch:

$$\hat{e}_c[n] = -be_c[n-T]$$

donde *T* es el periodo de pitch y *b* es la ganancia de predicción larga, y puede entenderse como el coeficiente LPC de predicción larga.

Al igual que en la predicción corta buscamos minimizar el error. Enunciaremos brevemente los pasos a seguir para el cálculo de *b* y *T*:

1. Definimos el error *J*.

$$J = \sum_n e[n]^2 = \sum_n (e_c[n] - \hat{e}_c[n])^2 = \sum_n (e_c[n] + be_c[n-T])^2$$

2. Minimizamos el error *J* respecto a la ganancia *b*.

$$\frac{\partial J}{\partial b} = 0 \rightarrow b = -\frac{\sum_n e_c[n]e_c[n-T]}{\sum_n e_c^2[n-T]}$$

Sustituyendo el valor de *b* el error queda solo en función de *T:*

$$J = \sum_n e_c^2[n] - \frac{\left(\sum_n e_c[n]e_c[n-T]\right)^2}{\sum_n e_c^2[n-T]}$$

3. Usaremos un algoritmo para calcular el periodo de pitch que minimiza el error. Para ello haremos un barrido del periodo en un intervalo donde la probabilidad de encontrar el pitch es alta.

1. $J_{min} = \infty$
2. *Para T desde* T_{min} *hasta* T_{max}
3. *Calcular b*
4. *Calcular J*
5. *Si* $J < J_{min}$
6. $J_{min} = J$
7. $b_{opt} = b$
8. $T_{opt} = T$
9. *Devolver* b_{opt} *y* T_{opt}

De esta manera hemos hallado el periodo de pitch óptimo para cada trama.

2.5.2. SÍNTESIS.

Durante el análisis hemos extraído una serie de parámetros, tales como los coeficientes LPC, la ganancia de predicción corta, el pitch y la ganancia de predicción larga. Con estos datos y el ruido blanco resultado del análisis podemos reconstruir la señal de voz.

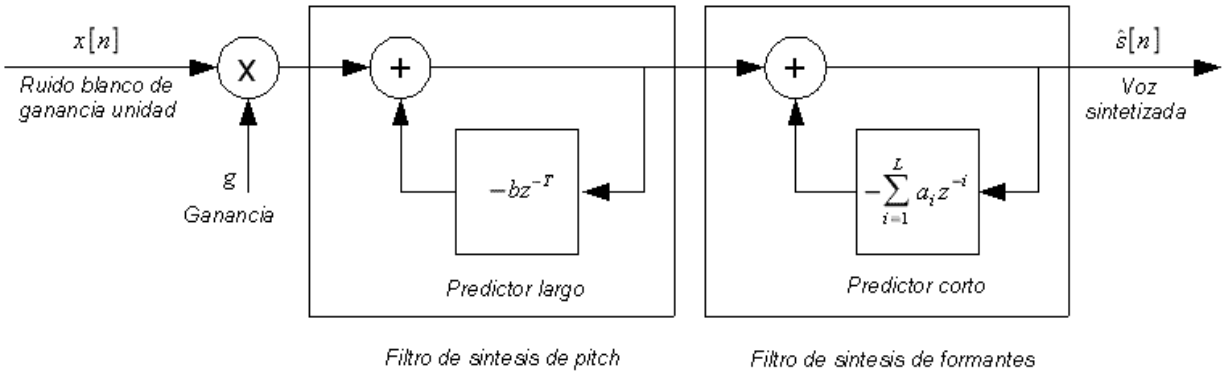

Figura 28: Modelo de producción de voz incluyendo predictor largo

El filtro de predicción larga es el responsable, en este caso, de generar correlación entre muestras separadas por un periodo de pitch, y presenta la siguiente función:

$$H(z) = \frac{1}{1 + bz^{-T}}$$

También se conoce como filtro de síntesis de pitch.

A continuación encontramos el filtro de predicción larga, que recrea la correlaciones entre muestras cercanas, con un orden comprendido entre 8 y 12, por lo general, 10. También se conoce como filtro de síntesis de formantes.

La ganancia *g* habitualmente puede obtenerse comparando la potencia de la señal sintetizada con la original.

2.6. CODIFICADOR CELP.

2.6.1. INTRODUCCIÓN.

CELP nació como un intento de mejora de LPC. Comparte con el algunos principios y difiere en otros aspectos:

1. Mantiene los modelos del predictor largo y corto.

2. Obvia la clasificación sonoro/sordo.

3. Introduce un libro de códigos que servirá de excitación para el proceso de síntesis.

2.6.2. MODELADO DEL PROCESO DE SÍNTESIS DE VOZ.

El proceso de síntesis de CELP se basa principalmente en los tres puntos anteriores. La excitación procede de un libro de códigos, que contiene un conjunto de excitaciones posibles. Estas excitación se escala mediante una ganancia y se hace pasar por los filtros de síntesis, sin tener en cuenta la distinción sonoro/sordo, es decir, se mantiene el mismo esquema para ambos tipos de sonido. Recordemos que en la práctica, en el esquema LPC se usaban distintas excitaciones para sintetizar sonidos sonoros/sordos.

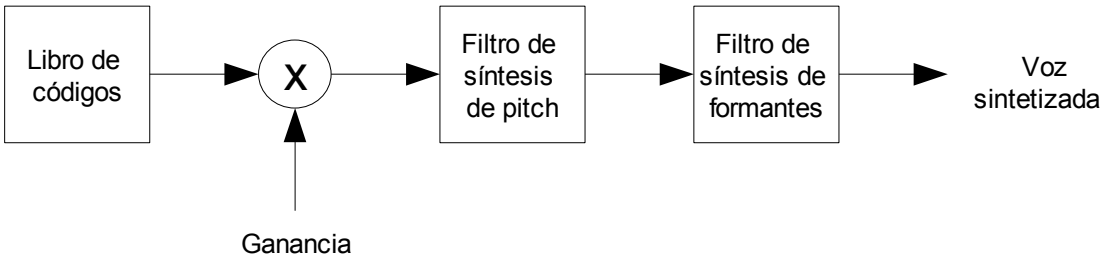

Figura 29: Modelo de producción de voz CELP

Los libros de código que contienen las excitaciones pueden ser fijos o adaptativos, y contener secuencias fijas o aleatorias. En este proyecto consideraremos un libro fijo con secuencias de ruido blanco.

Gracias a estos aspectos el modelo ofrece una mayor naturalidad en la voz sintetizada, eliminando las distinciones rígidas y manteniendo cierta información de fase, frente a LPC.

2.6.3. MODELADO DEL PROCESO DE ANÁLISIS DE VOZ.

Una vez introducido el proceso de síntesis, podemos tratar del análisis. CELP esta basado es una estructura "análisis mediante síntesis", un proceso de bucle cerrado, de tal forma que se extraen los parámetros, se sintetiza la voz y se compara con la muestra original, todo ello en el codificador.

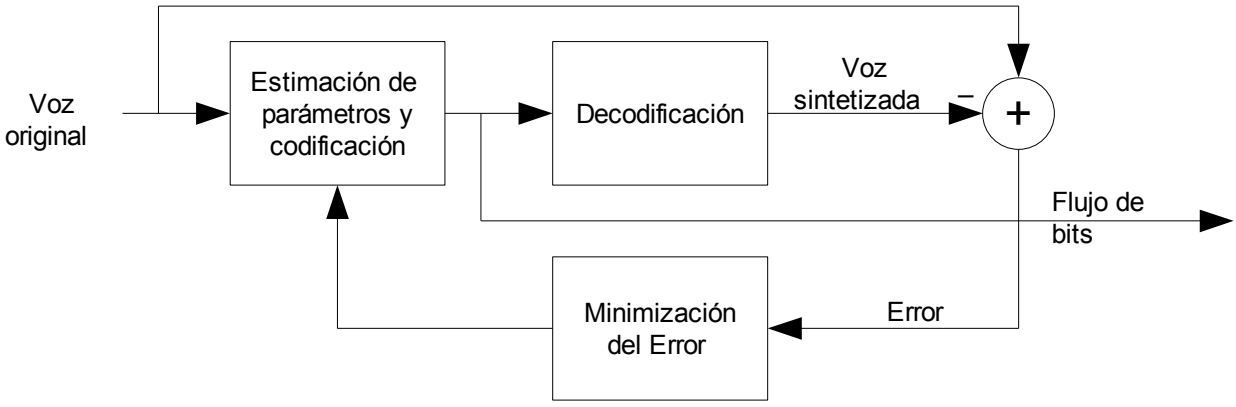

Figura 30: Codificador basado en el principio "análisis mediante síntesis"

El libro de códigos proporciona varias excitaciones, que iremos seleccionando mediante un índice. Para cada una de ellas calcularemos su correspondiente señal sintetizada, y nos quedaremos con los parámetros de aquella cuyo parecido con la fuente sea mayor.

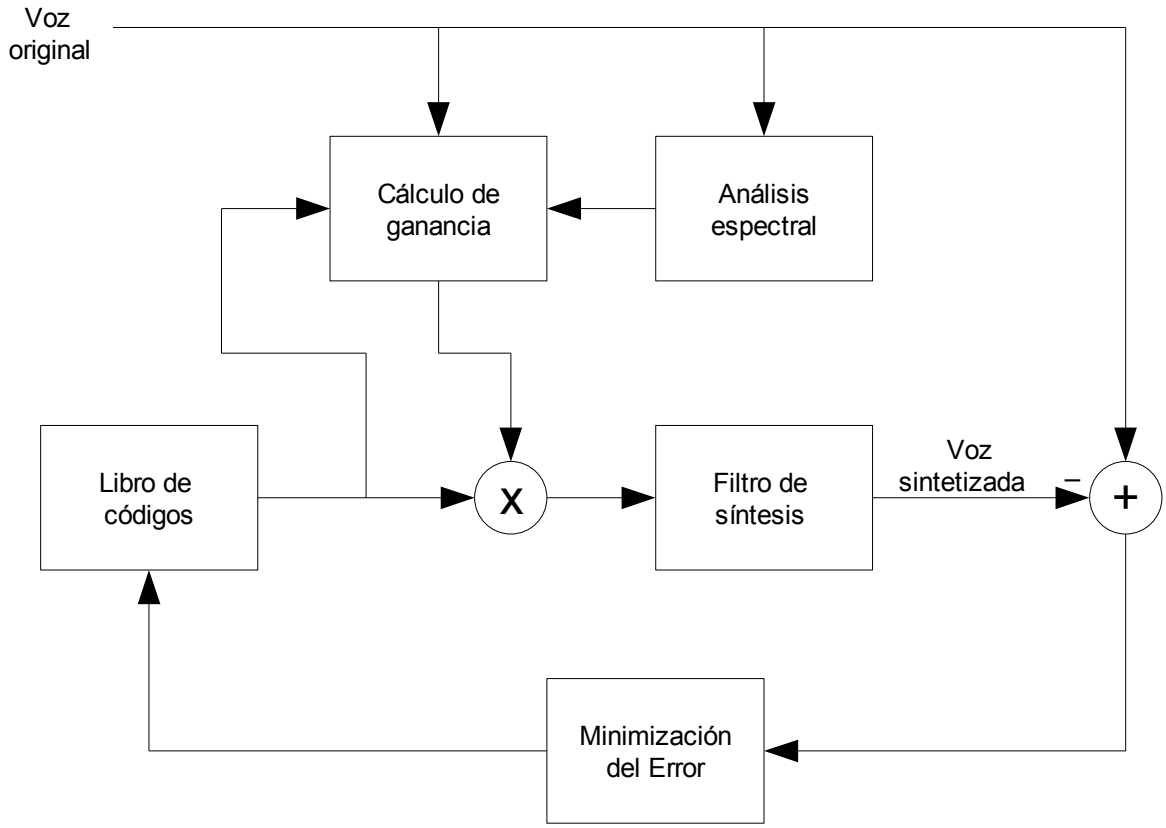

Figura 31: Componentes clave de un Codificador CELP

Desde el punto de vista computacional, el cálculo de todos los parámetros en bucle cerrado es costoso. Por ello los coeficientes de los filtros se calculan en bucle abierto. Una vez fijados estos, solo se varía el índice del libro de códigos, el único parámetro que se ajusta mediante realimentación.

2.6.4. CODIFICACIÓN Y DECODIFICACIÓN.

Filtro perceptual.

El bucle de "análisis mediante síntesis" en el que está basado el codificador CELP se muestra en la figura 32, donde la función de transferencia del filtro de síntesis de formantes viene dada por:

$$H_f(z)=\frac{1}{A(z)}=\frac{1}{1+\sum_{i=1}^{L} a_i z^{-i}}$$

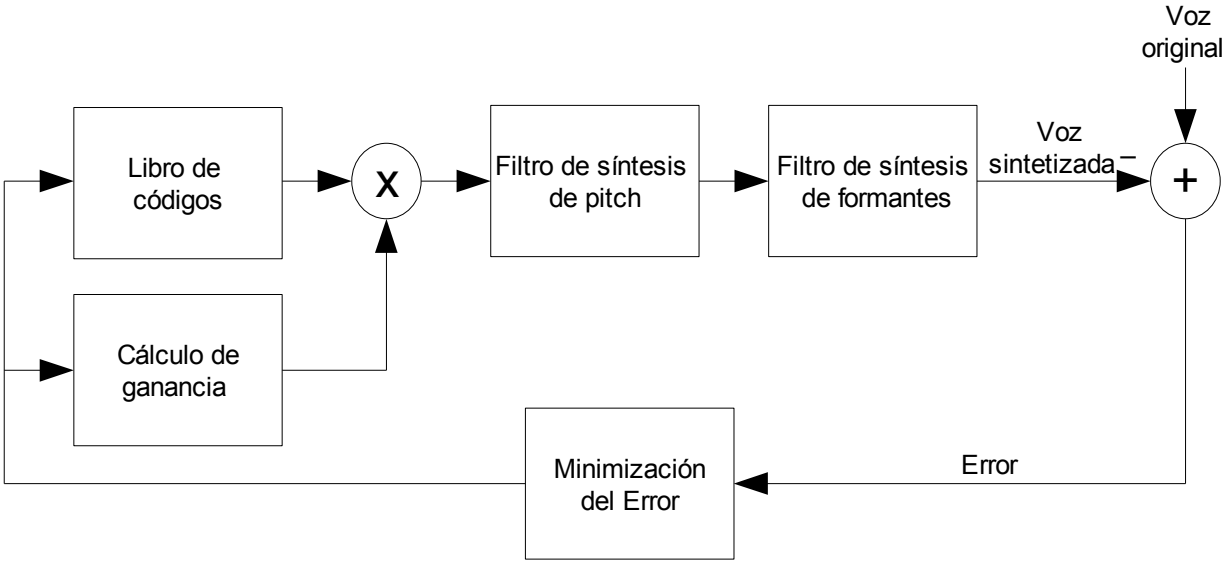

Figura 32: Bucle "análisis mediante síntesis" de un Codificador CELP

Es bien conocido que el oído humano es capaz de tolerar mayor cantidad de ruido en aquellas zonas en la que la señal de voz tiene más energía. Conociendo la estructura de los formantes, podemos "distribuir" el ruido de tal manera que quede enmascarado lo mayormente posible por la señal de voz. Para ello introduce un filtro perceptual antes de la etapa de minimización del error, que tiene la siguiente estructura:

$$W(z)=\frac{A(z)}{A(z/\gamma)}=\frac{1+\sum_{i=1}^{L} a_i z^{-i}}{1+\sum_{i=1}^{L} a_i \gamma^i z^{-i}}$$

con un valor de γ en el intervalo [0,1]. Su valor determina el grado en el que la señal de error decrece en todas las bandas. El filtro amplifica el error en las zonas en las que no hay formantes, y lo disminuye en las zonas en las que los hay.

Desde el punto de vista computacional, es preferible desplazar el filtro perceptual antes del sumador.

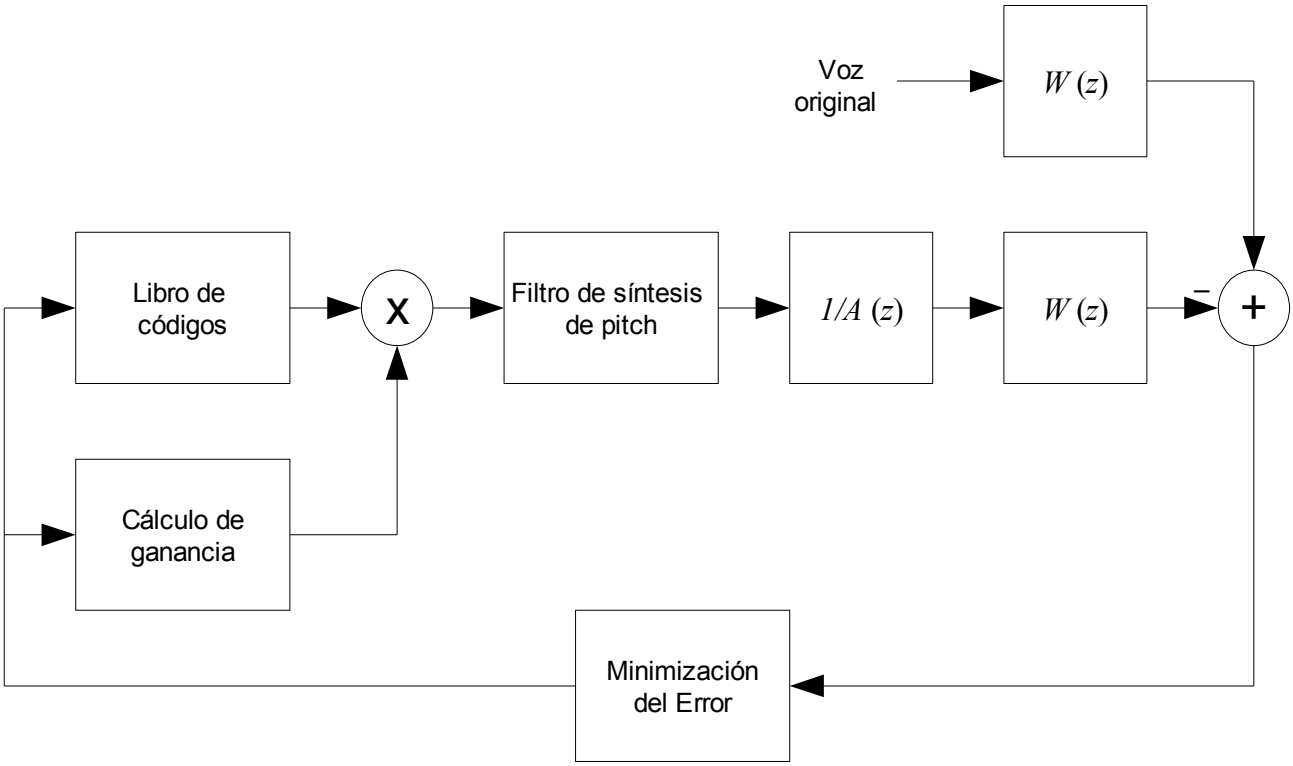

Figura 33: Bucle "análisis mediante síntesis" de un Codificador CELP con filtrado perceptual

De esta manera después del filtro de síntesis de formantes quedaría un filtro perceptual. Ambos pueden asociarse en un único sistema conocido como el filtro modificado de síntesis de formantes, con la siguiente estructura:

$$H_f(z/\gamma)=\frac{1}{A(z/\gamma)}=\frac{1}{1+\sum_{i=1}^{L} a_i \gamma^i z^{-i}}$$

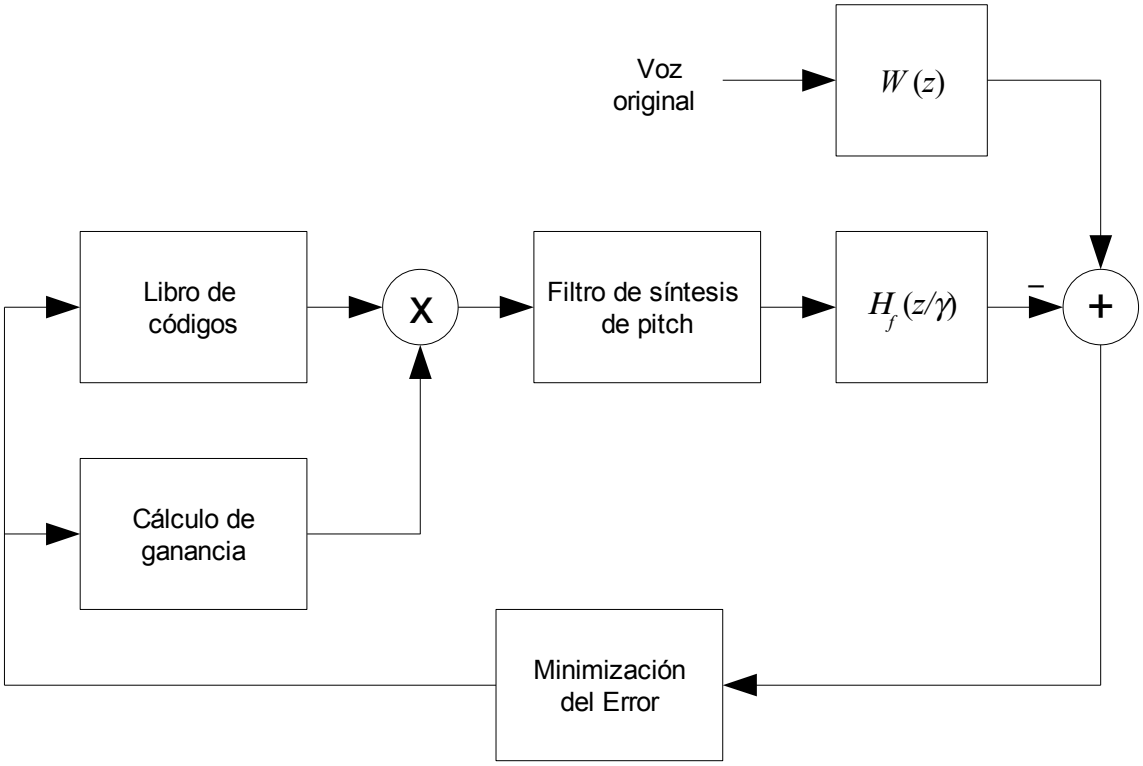

Figura 34: Bucle "análisis mediante síntesis" de un Codificador CELP con filtrado perceptual

Codificación.

El proceso de codificación consta a grosso modo de los siguientes pasos:

1. Segmentación. La señal de voz se dividirá en tramas de 20 a 30 milisegundos y subtramas de 5 a 7.5 milisegundos.

2. Análisis de coeficientes LPC para cada trama. Se usa como entrada la voz original.

3. Búsqueda de pitch y ganancia de predicción larga para cada subtrama. Habitualmente se usa como entrada el error de predicción corto.

4. Selección de secuencia de excitación para cada subtrama. Se recorre el libro de códigos, se sintetiza la voz para cada excitación, se compara con la original y se escoge el índice que minimice el error.

5. Transmisión de parámetros por cada subtrama. Se envía al decodificador el índice, la ganancia y los parámetros de los filtros de síntesis.

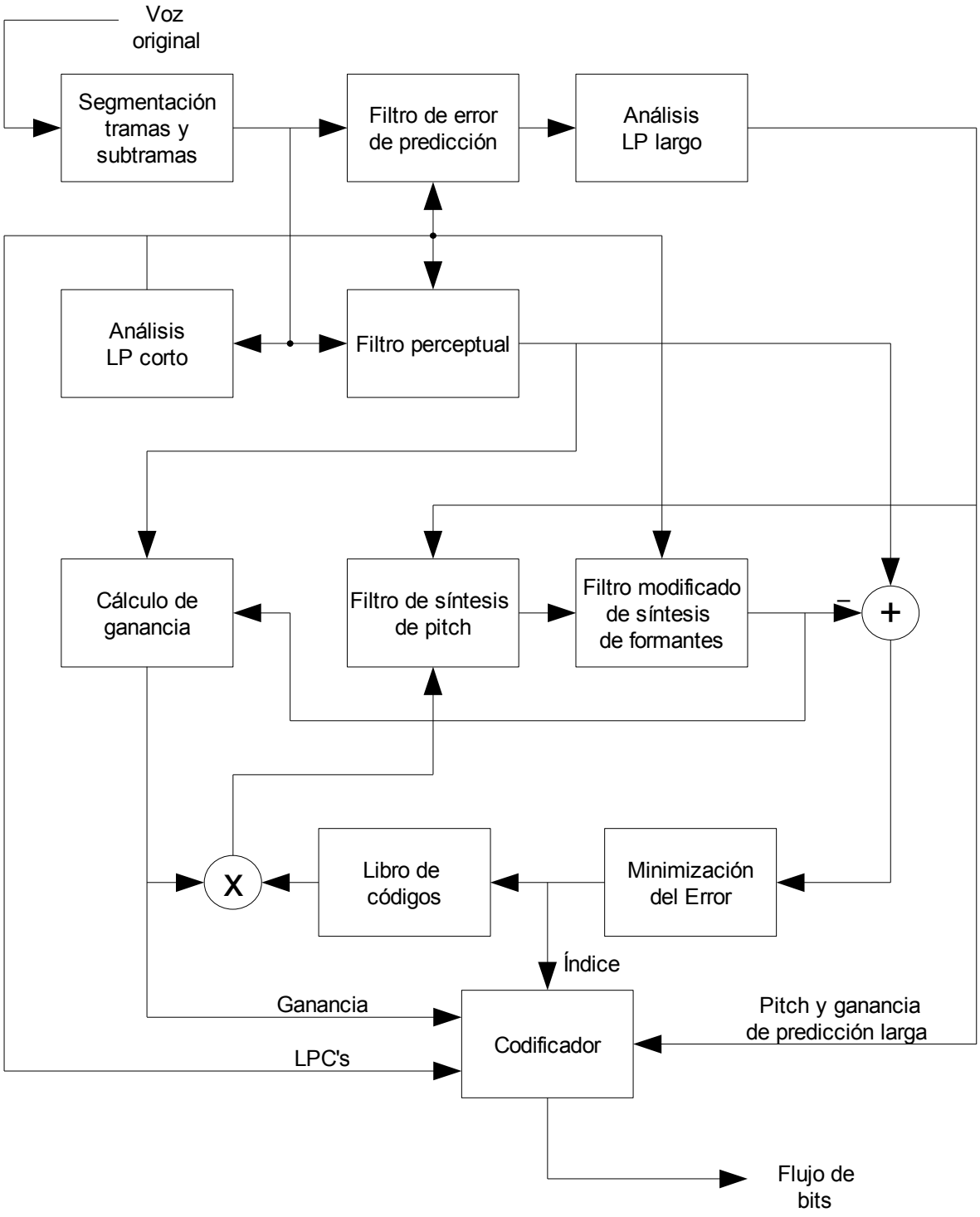

Figura 35: Codificador CELP genérico

Decodificación.

El proceso de decodificación consta a grosso modo de los siguientes pasos:

1. Recepción de parámetros por cada subtrama. Se recuperan los datos producto del análisis.

2. Síntesis. Con los parámetros recuperados fijamos la excitación, la ganancia y los filtros, y sintetizamos la subtrama.

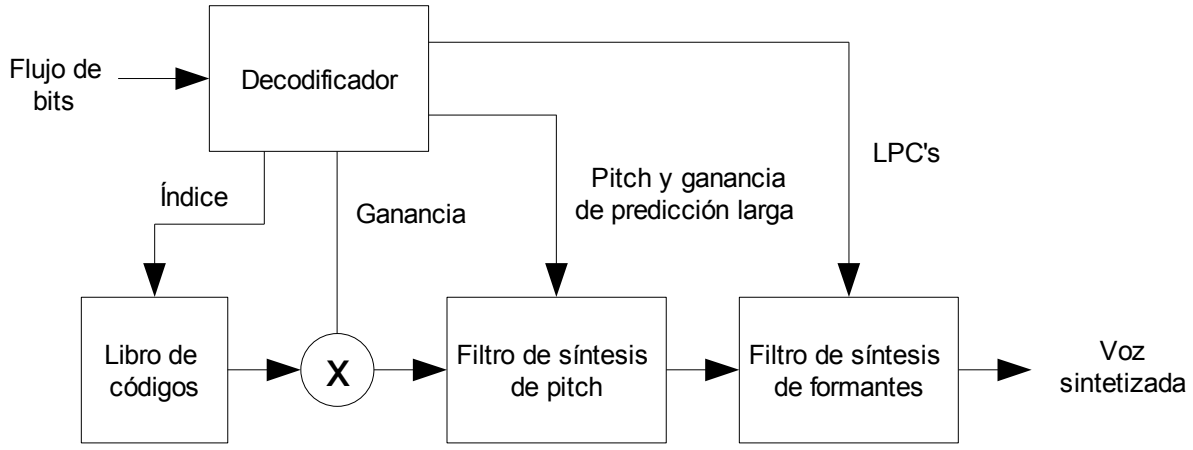

Figura 36: Decodificador CELP genérico

Búsqueda en el libro de códigos.

Se trata de la etapa más costosa desde el punto de vista computacional. Los pasos son:

1. Filtrado perceptual de la subtrama.

2. Para cada excitación del libro de códigos:

   2.1. Cálculo de la ganancia.

2.2. Escalado de la excitación.

2.3. Paso por el filtro de síntesis de pitch de la señal del paso 2.2.

2.4. Paso por el filtro modificado de síntesis de formantes de la señal de salida del paso 2.3, obteniéndose la señal de voz sintetizada.

2.5. Obtención del error como diferencia entre la señal original y la señal sintetizada.

2.6. Cálculo de la energía de la señal de error.

2.7. Retención del índice de la excitación que da lugar al error con menor energía.

Cálculo de la ganancia.

Para realizar estos cálculos consideraremos la descomposición del los filtros en cascada es sus respectivas respuestas ante estado-cero y entrada-cero, y después introduciremos este modelo en el bucle cerrado.

Estado-cero y entrada cero.

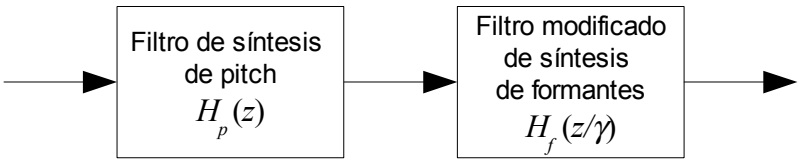

Figura 37: Conexión de un filtro de síntesis de pitch y un filtro de síntesis de formantes modificado

Consideremos el sistema de la figura 37 con sus respectivas ecuaciones en diferencias:

$$y[n]=d[n]-\sum_{i=1}^{L} a_i \gamma^i y[n-i] \quad , \quad d[n]=x[n]-bd[n-T]$$

donde

- L es el orden del predictor corto.
- a_i representan los coeficientes LPC.
- b es la ganancia de predicción larga.
- T es el periodo de pitch.

En la siguiente figura mostramos la estructura del sistema descompuesto en estado-cero y respuesta cero:

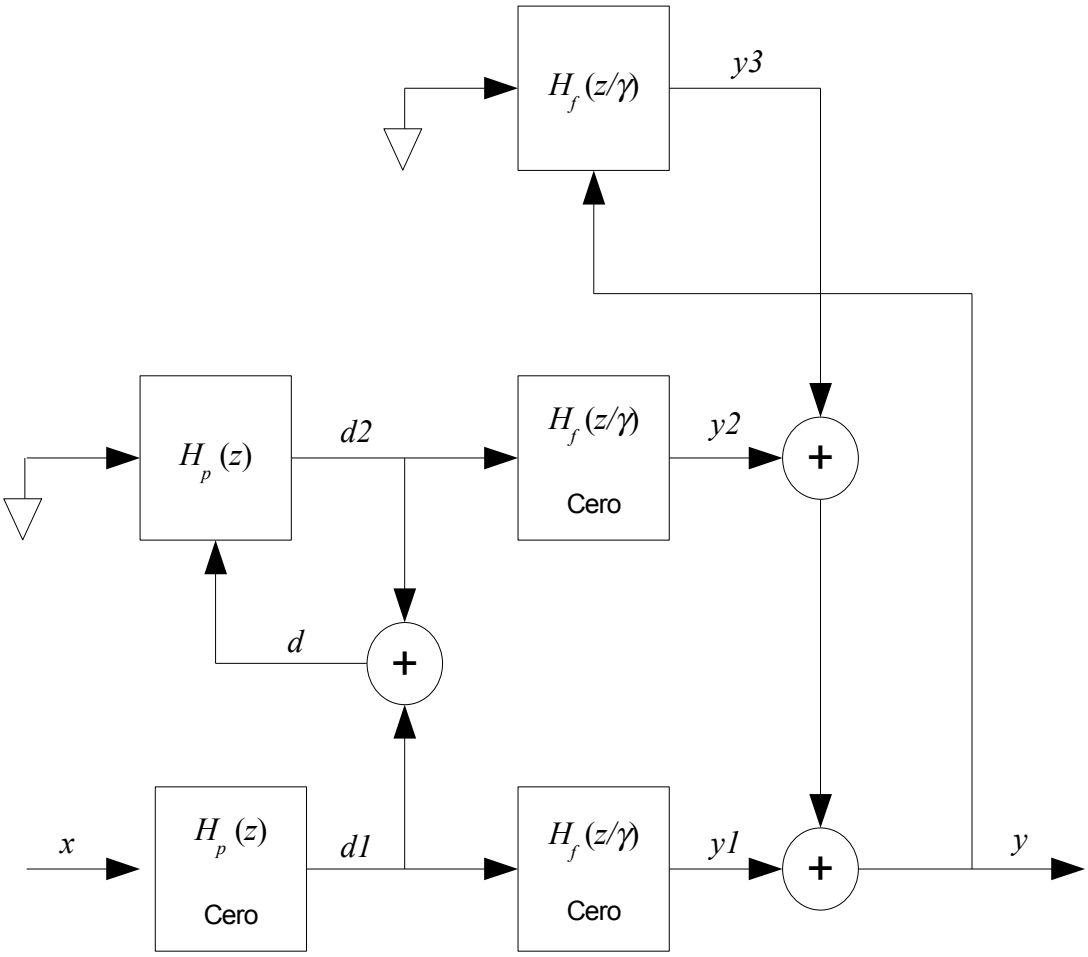

Figura 38: Señales involucradas en el método entrada-cero estado-cero

A continuación nos centraremos en el cálculo de las respuestas:

- Cálculo de *d[n]*

1. Respuesta ante estado-cero:

$$d1_r[n]=x_r[n] \qquad 0\le n\le T-1$$

$$d1_r[n]=x_r[n]-bd1_r[n-T] \qquad T\le n\le N-1$$

2. Respuesta ante entrada-cero:

$$d2_r[n]=d_{r-1}[n+N] \qquad -T\le n\le -1$$

$$d2_r[n]=-bd2_r[n-T] \qquad 0\le n\le N-1$$

3. Respuesta conjunta:

$$d[n]=d1_r[n]+d2_r[n] \qquad 0\le n\le N-1$$

- Cálculo de *y[n]*

1. Respuesta ante estado-cero:

$$y1_r[n]=0 \qquad -L\le n\le -1$$

$$y1_r[n]=d1_r[n]-\sum_{i=1}^{L} a_i\gamma^i y1_r[n-i] \qquad 0\le n\le N-1$$

2. Respuesta ante entrada-cero del filtro de síntesis de pitch:

$$y2_r[n]=0 \qquad -L\le n\le -1$$

$$y2_r[n]=d2_r[n]-\sum_{i=1}^{L} a_i\gamma^i y2_r[n-i] \qquad 0\le n\le N-1$$

3. Respuesta ante entrada-cero del filtro de síntesis de formantes:

$$y3_r[n]=y_{r-1}[n+N] \qquad -L\le n\le -1$$

$$y3_r[n]=-\sum_{i=1}^{L} a_i\gamma^i y3_r[n-i] \qquad 0\le n\le N-1$$

4. Respuesta conjunta:

$$y_r[n]=y1_r[n]+y2_r[n]+y3_r[n] \qquad 0\le n\le N-1$$

Cálculo del error y la ganancia óptima.

En el apartado anterior hemos estudiado la descomposición de la respuesta de los filtros en estado-cero y respuesta-cero. Introduciremos el modelo en el esquema de búsqueda en el libro de códigos. Nuestro objetivo en este apartado es elaborar una expresión del error dependiente de la ganancia, y minimizarlo respecto a ese parámetro.

La estructura a analizar sería la siguiente:

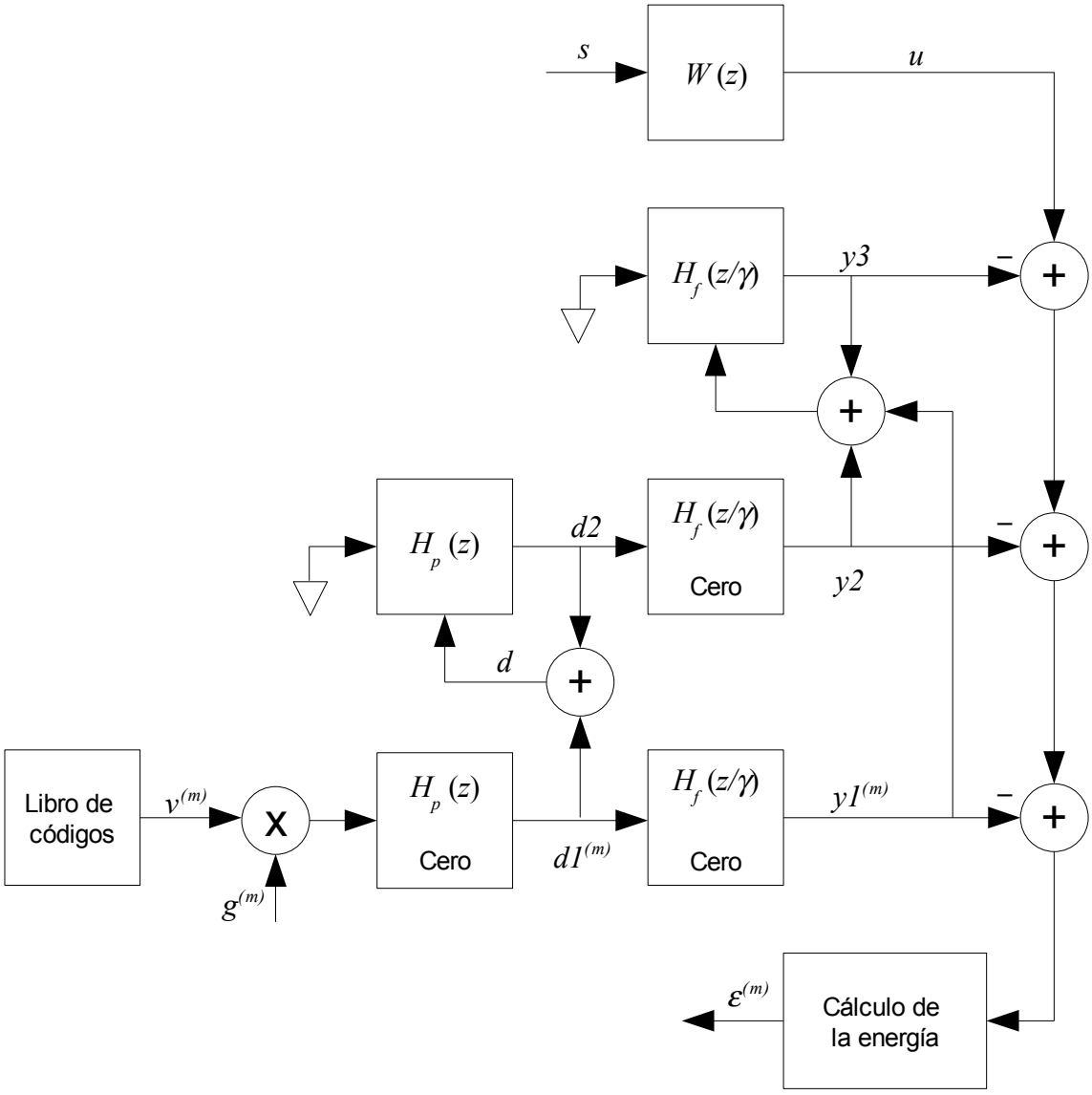

Figura 39: Señales involucradas en la búsqueda del libro de códigos

El libro de códigos es fijo y contiene M vectores, denotados por:

$$v^{(m)}[n] \qquad m=0,1,\ldots,M-1 \qquad n=0,1,\ldots,N-1$$

Estos vectores son escalados mediante la ganancia *g(m)*, diferente para uno de ellos. Estos vectores escalados serán las entradas de los filtros.

El error tiene varias componentes:

- Una componente resultado del paso de la subtrama por el filtro perceptual, la señal *u*.
- Las respuestas ante entrada-cero, *y2* e *y3*.
- La respuesta ante estado-cero, *y1(m)*.

La energía del error viene dada por la siguiente expresión:

$$\varepsilon^{(m)}=\sum_{n=0}^{N-1}(u[n]-y1^{(m)}[n]-y2[n]-y3[n])^2$$

Introducimos la respuesta ante estado-cero sin escalar:

$$y1_o^{(m)}[n]=y1^{(m)}[n]/g^{(m)} \qquad n=0,\ldots,N-1$$

y agrupamos el resto de componentes del error:

$$u_o[n]=u[n]-y2[n]-y3[n]$$

quedando su energía:

$$\varepsilon^{(m)}=\sum_{n=0}^{N-1}(u_o[n]-g^{(m)}y1_o^{(m)}[n])^2$$

Minimizando la energía del error respecto a la ganancia se obtiene:

$$g^{(m)}=\frac{\sum_{n=0}^{N-1}u_o[n]y1_o^{(m)}[n]}{\sum_{n=0}^{N-1}(y1_o^{(m)}[n])^2}$$