

Codificación de voz

EL 7022 – Procesamiento de voz en telecomunicaciones

Prof. Néstor Becerra Yoma, Ph.D.

Laboratorio de Procesamiento y Transmisión de Voz
Universidad de Chile



Universidad
de Chile



DEPARTAMENTO DE
INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS
FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Contenidos

- Modelo de producción de la voz
- Codificación de la voz
- Vocoder LPC
- Análisis por síntesis
- Vocoder MELP (Introducción y estándar MIL-STD-3005)

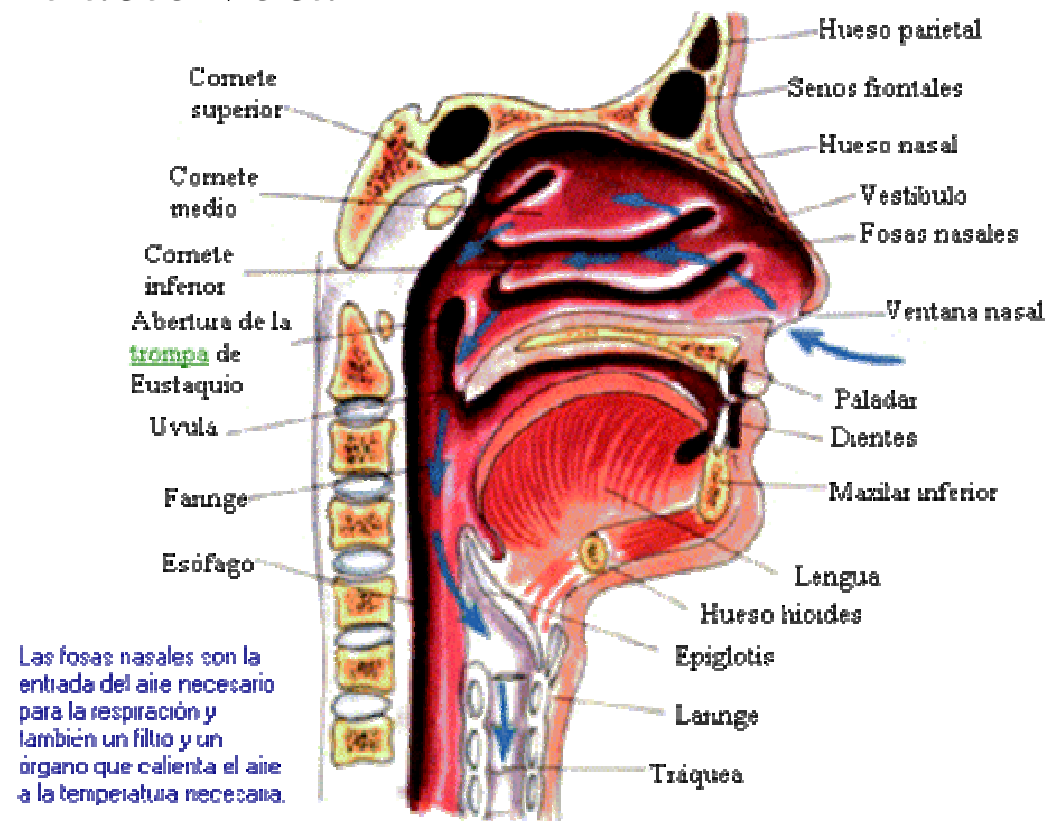
Modelo de producción y de percepción de la señal de voz

- Filtro del tracto vocal
- La frecuencia fundamental y las cuerdas vocales
- Formantes



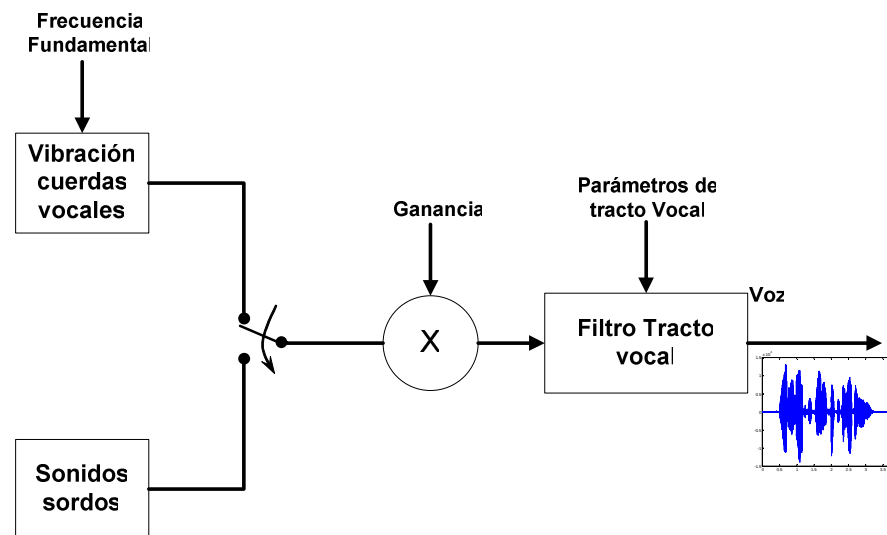
Modelo de producción y de percepción de la señal de voz

□ Filtro del tracto vocal



Modelo de producción y de percepción de la señal de voz

- Filtro del tracto vocal
 - ▣ Modelo del filtro del tracto vocal

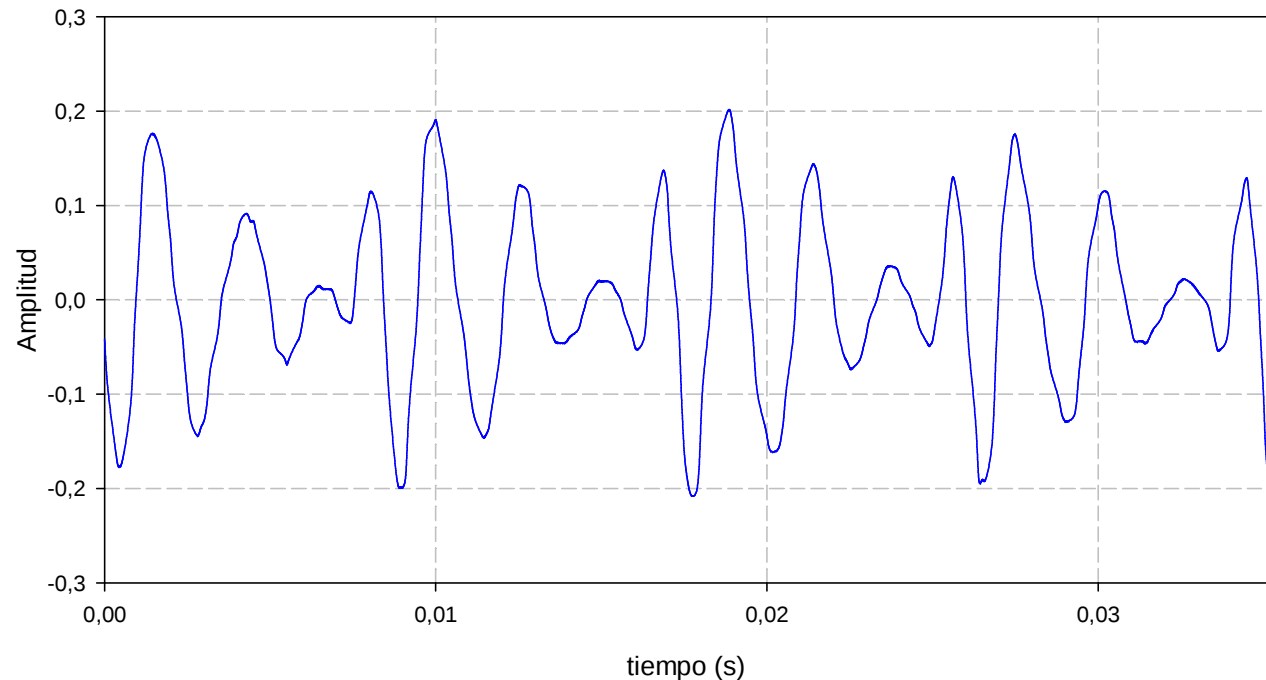


Modelo de producción y de percepción de la señal de voz

- La frecuencia fundamental y las cuerdas vocales
 - ▣ Vibración de las cuerdas vocales
 - ▣ La frecuencia fundamental (FO) corresponde al tono
 - ▣ La percepción del tono se conoce como Pitch

Modelo de producción y de percepción de la señal de voz

- La frecuencia fundamental y las cuerdas vocales



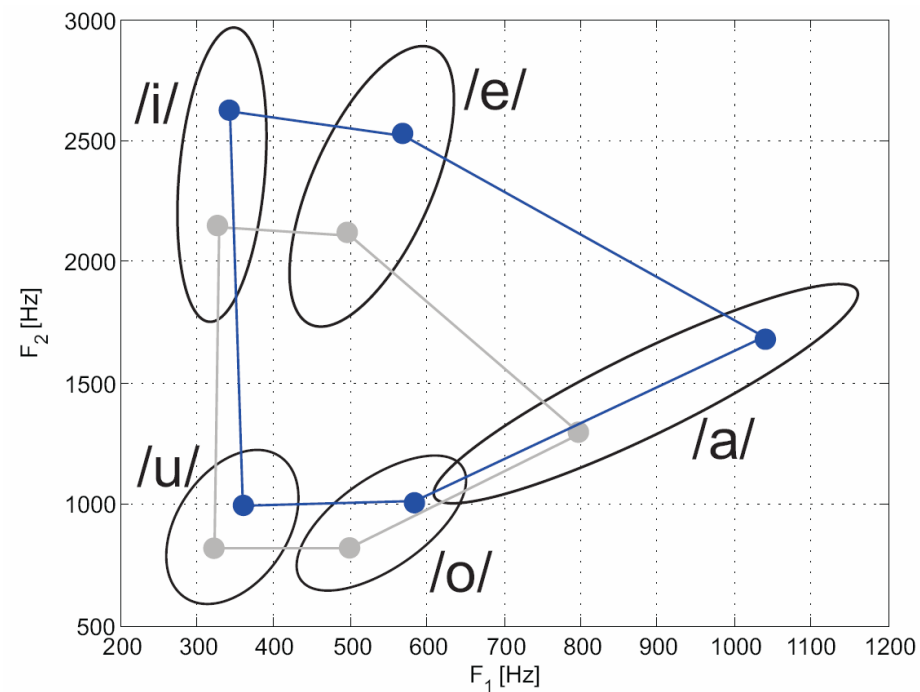
Modelo de producción y de percepción de la señal de voz

□ Formantes

- Los formantes (de forma) son zonas en el espectro de frecuencias en las que todo armónico que cae en esas zonas es amplificado y potenciado.
 - Polos del LPC
 - Concentraciones de energía

Modelo de producción y de percepción de la señal de voz

□ Formantes



Codificación

- ¿Qué es codificación?
- Tipos de codificadores de voz
 - ▣ Forma de onda
 - ▣ Vocoders

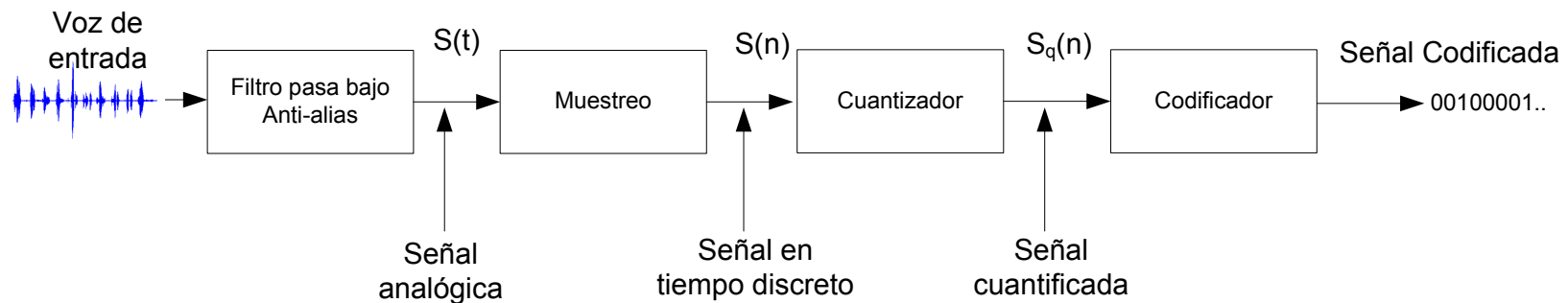


Codificación

- ¿Qué es codificación?
 - ▣ Comprimir información para transmitir o almacenar
 - ▣ Codificadores.
 - Para enviar.
 - ▣ Decodificadores.
 - Para Recibir.
 - ▣ Sistema de codificación y decodificación se conoce como “CODEC”.

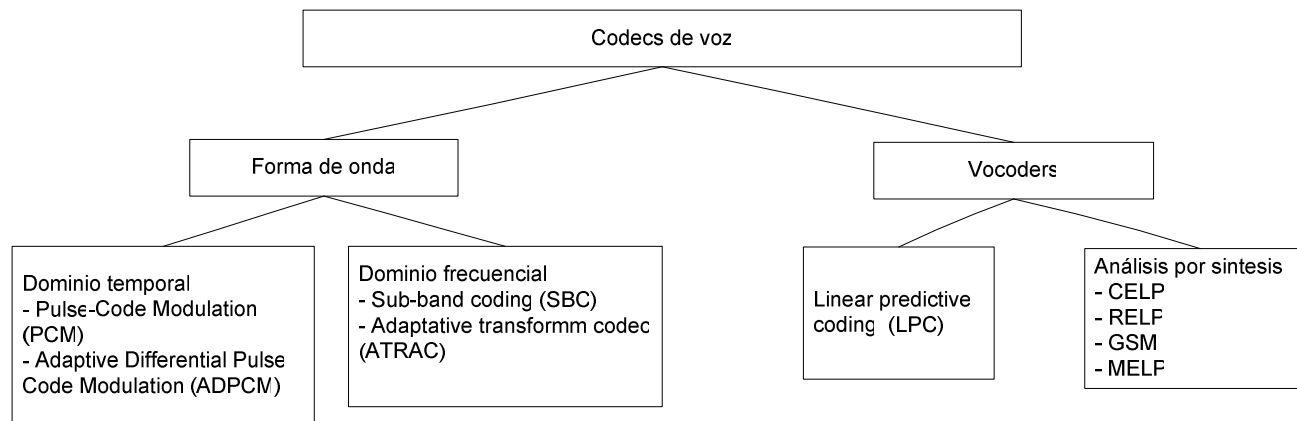
Codificación

- ¿Qué es codificación?
 - ▣ Esquema Básico de codificación:



Codificación

- Tipos de codificadores de voz
 - ▣ Forma de Onda: Preservan la forma de onda de la señal
 - ▣ Vocoders: Analizan la señal y extraen los parámetros del modelo de producción de la voz



Codificación

- Tipos de codificadores de voz
 - ▣ Forma de Onda
 - Señal decodificada semejante a original
 - Utiliza características de la voz en mayor o menor grado
 - En términos generales la señal de voz es filtrada, muestreada, cuantizada y codificada.

Codificación

- Tipos de codificadores de voz
 - ▣ Forma de Onda
 - Los más conocidos:
 - Pulse Code Modulation (PCM) – ITU G711
 - Ley A
 - Ley μ
 - Adaptive Delta Pulse Code Modulation (ADPCM) – ITU G726

Codificación

□ Tipos de codificadores de voz

■ Vocoders

- El propósito es poder transmitir voz maximizando la calidad subjetiva con el menor ancho de banda posible
- Modelos basados en la producción y recepción de la voz.
- Codificación por frames, no muestra a muestra
- Codificadores con tasas incluso por debajo de 2.5 kbps.

Codificación

□ Tipos de codificadores de voz

□ Vocoders

■ Vocoder conocidos:

■ Básico

■ LPC

■ Análisis por síntesis

■ CELP

■ RELP

■ GSM

■ Análisis por síntesis + Excitación mezclada

■ MELP

Codificación

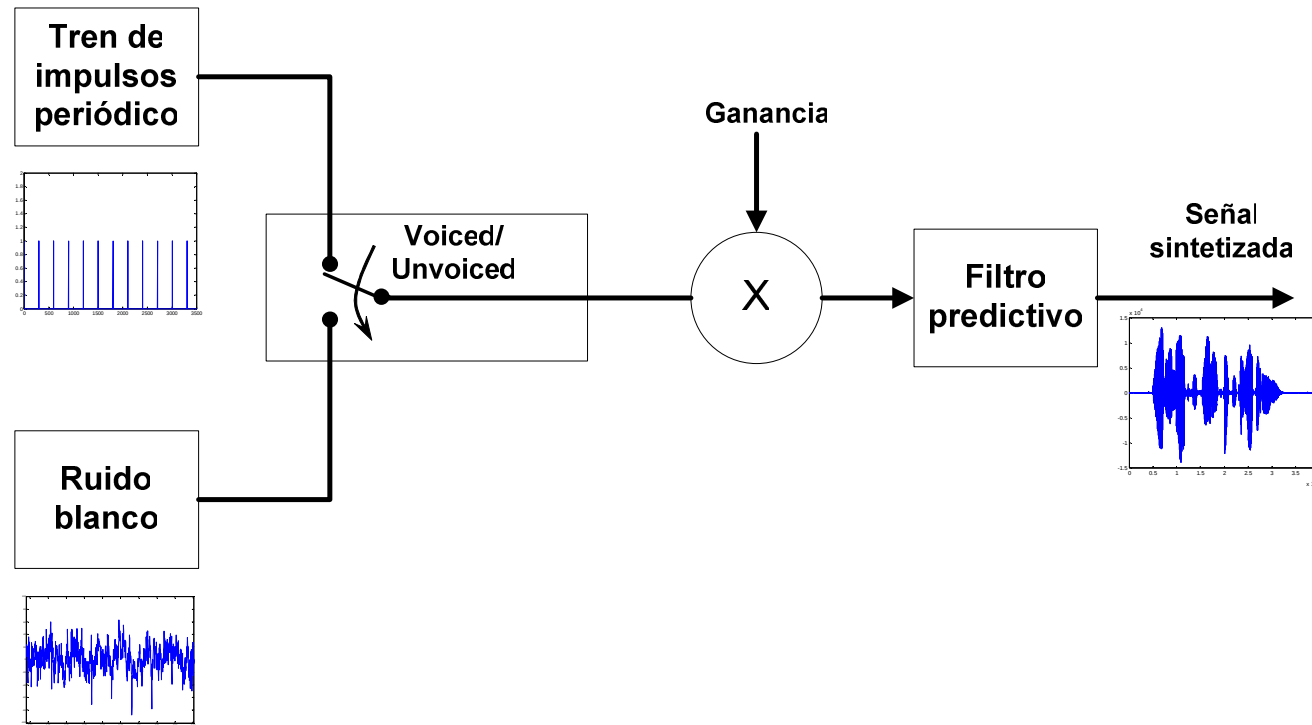
- Medidas de calidad de la voz en codificación
 - ▣ Métodos subjetivos
 - La calidad de la voz se establece a través de la opinión del usuario
 - Mean Opinión Store (MOS)
 - ▣ Métodos objetivos
 - Comparación de una señal inicial con una señal degradada
 - Perceptual Evaluation of Speech Quality (PESQ)

LPC (Linear Predictive Coding)

- Modelo predictivo lineal del conducto vocal
- La idea principal se basa en que la señal de voz puede modelarse a través de una combinación lineal de p muestras anteriores más una señal de excitación o ruido blanco.

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

LPC (Linear Predictive Coding)



LPC (Linear Predictive Coding)

- Para determinar los coeficientes del LPC se pueden deducir las ecuaciones de Yule-Walker a partir de la minimización del error cuadrático medio

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left[s(n) - \sum_{k=1}^p \alpha_k s(n-k) \right]^2$$

$$\frac{\partial E}{\partial \alpha_k} = 0 \quad k = 1, 2, 3, \dots, p$$

LPC (Linear Predictive Coding)

- Luego se tiene p ecuaciones lineales:

$$\sum_{n=-\infty}^{\infty} s(n-i)s(n) = \sum_{k=1}^p \alpha_k \sum_{n=-\infty}^{\infty} s(n-i)s(n-k) \quad i = 1, 2, 3, \dots, p$$

- Si definimos:

$$R(i) = \sum_{n=i}^{N-1} s(n)s(n-i) \quad i = 1, 2, 3, \dots, p$$

- Reemplazando:

$$\sum_{k=1}^p \alpha_k R(i-k) = R(i) \quad (*) \quad i = 1, 2, 3, \dots, p$$

LPC (Linear Predictive Coding)

- De las ecuaciones anteriores tenemos que, la energía residual mínima o el error de predicción mínima E_p para un modelo de p polos

$$E_p = R(0) - \sum_{k=1}^p \alpha_k R(k)$$

Ecuación (*) se puede representar matricialmente como:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ \vdots \\ R(p) \end{bmatrix}$$

LPC (Linear Predictive Coding)

- Resumiendo:

$$R * A = r$$

define las ecuaciones de Yule-Walker y donde R es una matriz de $p \times p$. Invirtiendo la matriz R se puede obtener el vector A de coeficientes del LPC.

- Método de Levinson-Durbin

Consiste en aprovechar las simetrías de la matriz.

LPC (Linear Predictive Coding)

□ Método de Levinson-Durbin

$$k_m = \frac{R(m) - \sum_{i=1}^{m-1} \alpha_i^{m-1} R(m-i)}{E_{m-1}}$$

$$\alpha_m^m = k_m$$

$$\alpha_i^m = \alpha_i^{m-1} - k_m \alpha_{m-i}^{m-1} \quad 1 \leq i \leq m-1$$

LPC (Linear Predictive Coding)

□ Método de Levinson-Durbin

$$E_m = (1 - k_m^2) \cdot E_{m-1}$$

Donde inicialmente:

$$E_0 = R(0)$$

$$\alpha_0 = 0$$

Donde α_i^m indicando el coeficiente LPC i en la m -ésima iteración

Cuantización vectorial (VQ)

□ Motivación:

LPC $\sim$ 10 coeficientes (10 float 4 bytes)

Pitch (float 4 bytes)

V/U (1 bits)

Ganancia (float 4 bytes)

$$4 \text{ bytes} \times 8 \frac{\text{bits}}{\text{bytes}} \times 10 + 1 \text{ bytes} \times 8 \frac{\text{bits}}{\text{bytes}} + 1 \text{ bytes} \times 8 \frac{\text{bits}}{\text{bytes}} + 1 \text{ bits} = 337 \text{ bits}$$

$$337 \text{ bits} \times 80 \frac{\text{frames}}{\text{s}} \Rightarrow 26.96 \text{ kbps}$$

Usando cuantización vectorial $\sim$ 10 coeficientes (Codebook 5 bits – 32 celdas)

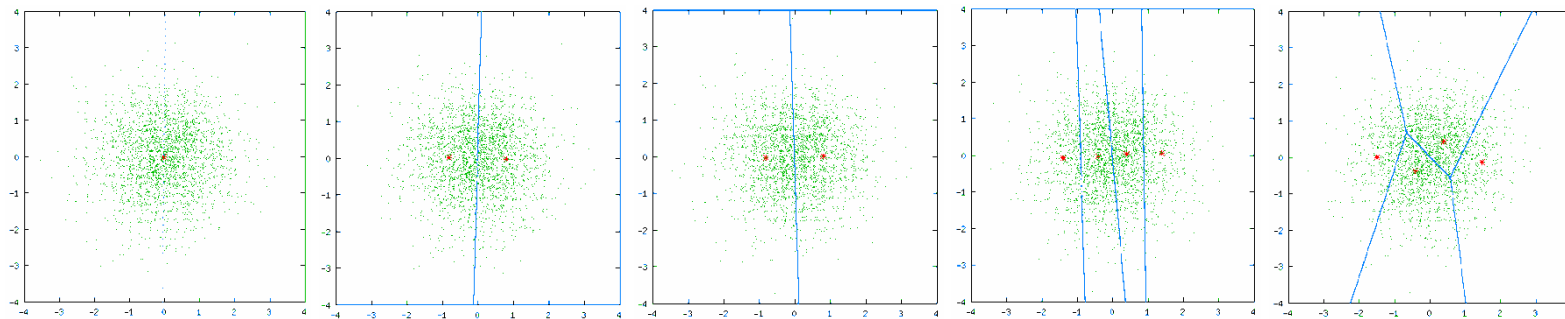
Pitch y ganancia (4 bits); V/U (1 bits)

$$5 \text{ bits} + 4 \text{ bits} + 4 \text{ bits} + 1 \text{ bits} = 14 \text{ bits}$$

$$14 \text{ bits} \times 80 \frac{\text{frames}}{\text{s}} \Rightarrow 1.12 \text{ kbps}$$

Cuantización vectorial (VQ)

- La cuantización es una herramienta muy útil para codificación pues el número de bits a representar es una variable que puede ser reducida cuanto se quiera cumpliendo una exigencia de distorsión máxima exigida.



Cuantización vectorial (VQ)

- Objetivo:

- ▣ Obtener un pequeño conjunto de vectores (**codebook**) que representen al espacio de muestras de forma que cada bloque se pueda representar por el índice de su vector asociado (**codeword**). Cada bloque puede ser representado por su codeword (centroide) o puede ser calculado como una densidad de función de probabilidad.

Cuantización vectorial (VQ)

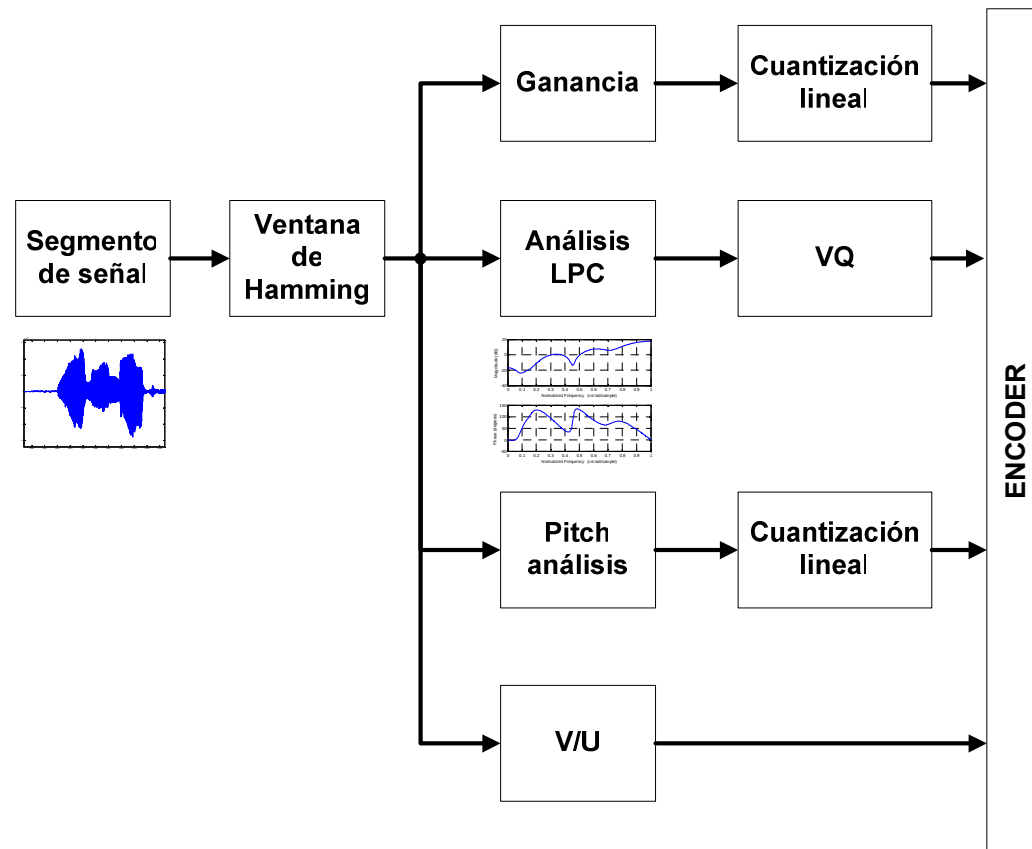
- Un método comúnmente utilizado en la elaboración de un codebook es el algoritmo “**k-means**”. Este método emplea una minimización de la distorsión global, definida como:

$$D = \sum_{i=1}^L D_i$$

donde D_i es la distorsión dentro de la celda i .

Vocoder LPC

■ Encoder



Análisis por Síntesis

Motivación

- A pesar de que la voz sintetizada por el Vocoder LPC-10 es inteligible, es posible observar los siguientes problemas:
 - El uso de señales de excitación artificiales produce un sonido sintético y “robotizado”.
 - Es difícil identificar al locutor.
 - La voz codificada no es analizada para chequear eficiencia (no existe control de la distorsión por codificación).

Análisis por Síntesis

- Vocoders de Predicción Lineal Basados en Análisis por Síntesis:
 - Uso de análisis de lazo cerrado para determinar la secuencia de excitación.
 - Se utiliza la secuencia de excitación que minimiza la diferencia entre las señales de voz de entrada y la sintetizada.

Análisis por Síntesis

- Error de síntesis o “residuo”

- ▣ Recuperar la señal original $s[n]$ agregando el residuo de la síntesis, $e[n]$:

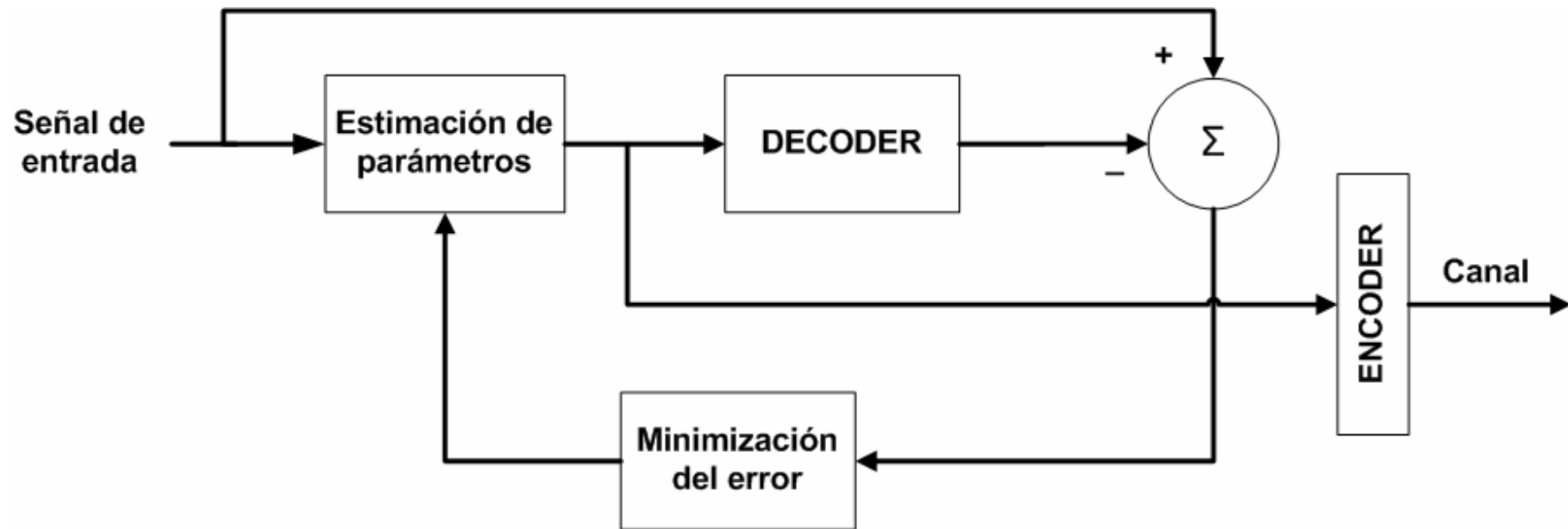
$$s[n] = \hat{s}[n] + e[n]$$

- ▣ Si el decoder recibe $e[n]$ en paralelo a los parámetros de la síntesis, es posible recuperar $s[n]$.

Análisis por Síntesis

- Predicción de lazo cerrado
 - Para asegurarse que el encoder y el decoder se mantendrán “sincronizados” **se implementa el decoder en el encoder.**
 - Es decir, se envía información complementaria asociada al error de predicción.
 - De esta forma, la señal que ingresa al encoder será idéntica a la generada en el decoder, no se acumulan diferencias.

Análisis por Síntesis



Análisis por Síntesis

La pregunta ahora es: **¿Cómo comprimir el residuo?**

Análisis por Síntesis

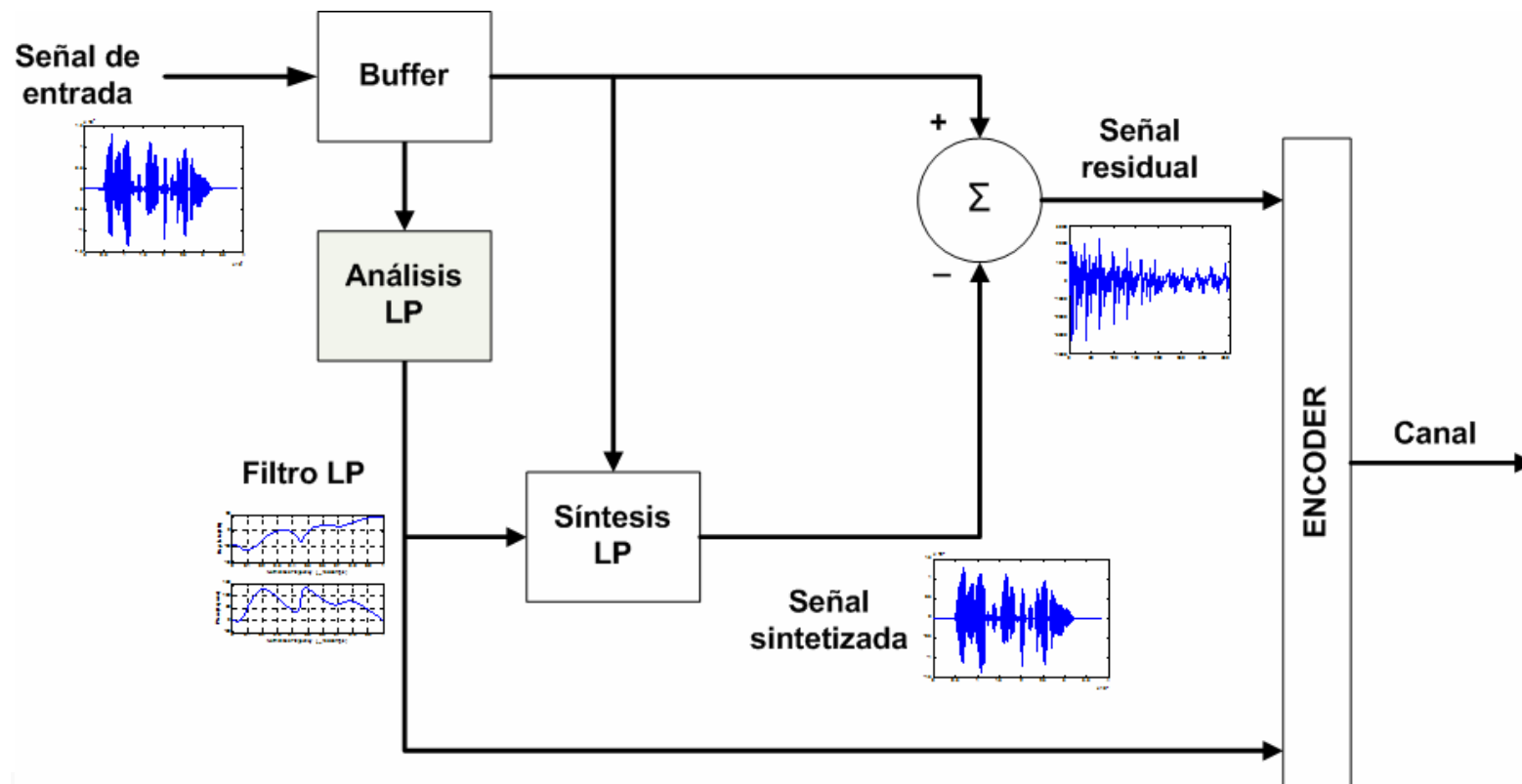
- Vocoders que utilizan el método de Análisis por Síntesis
 - ▣ RELP
 - ▣ RPE-LPT (GSM)
 - ▣ CELP
 - ▣ MELP

RELP

- Residual-Excited Linear Predictive Vocoder
 - Se realiza análisis LPC-10
 - Utiliza la señal residual, $e[n]$, como señal de excitación en el proceso de síntesis.
 - Codifica la señal residual para su transmisión.

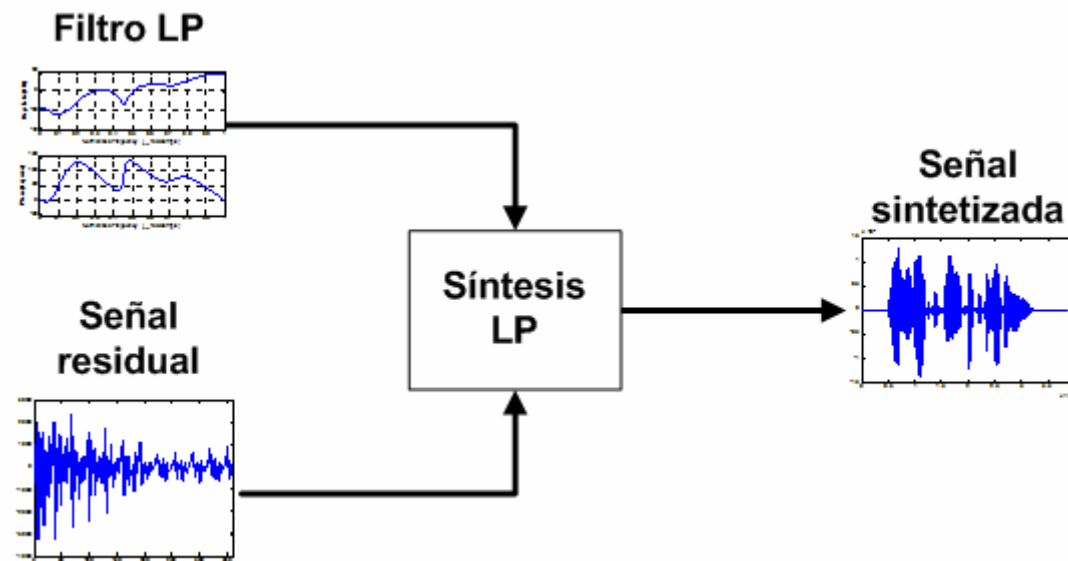
RELAP

□ Diagrama de Bloques - Encoder



RELAP

□ Diagrama de Bloques - Decoder

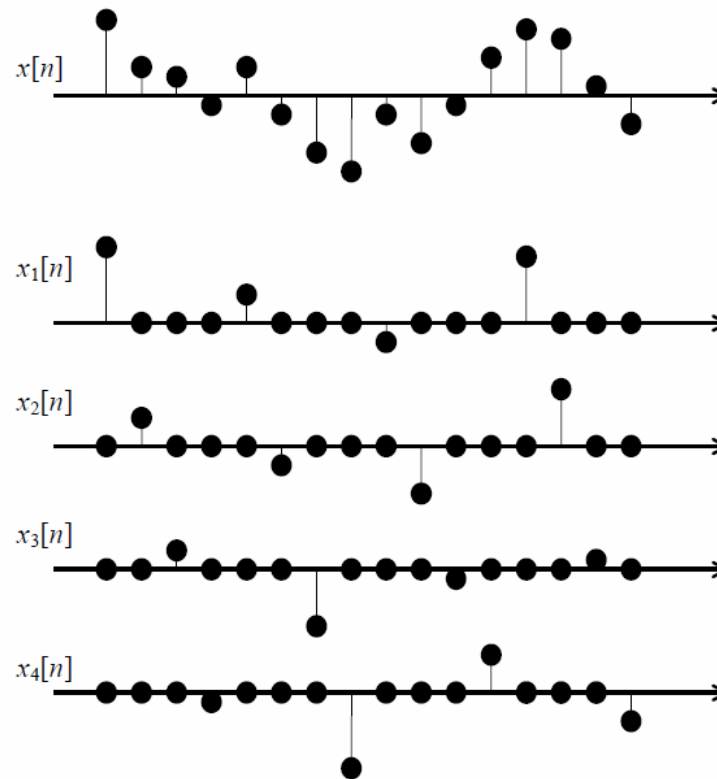


RPE-LPT (GSM)

- Regular Pulse Excited - Long Term Prediction Vocoder
 - Utilizado en el estándar celular **GSM**
 - Short term predictor: se realiza análisis LPC-10
 - **Long term predictor:** se realiza análisis de Pitch y Ganancia
 - **Regular Pulse Excited:** la señal residual, $e[n]$, se modela como una suma ponderada de 4 secuencias de pulsos.
 - El residuo se codifica con el índice y la amplitud de cada secuencia.

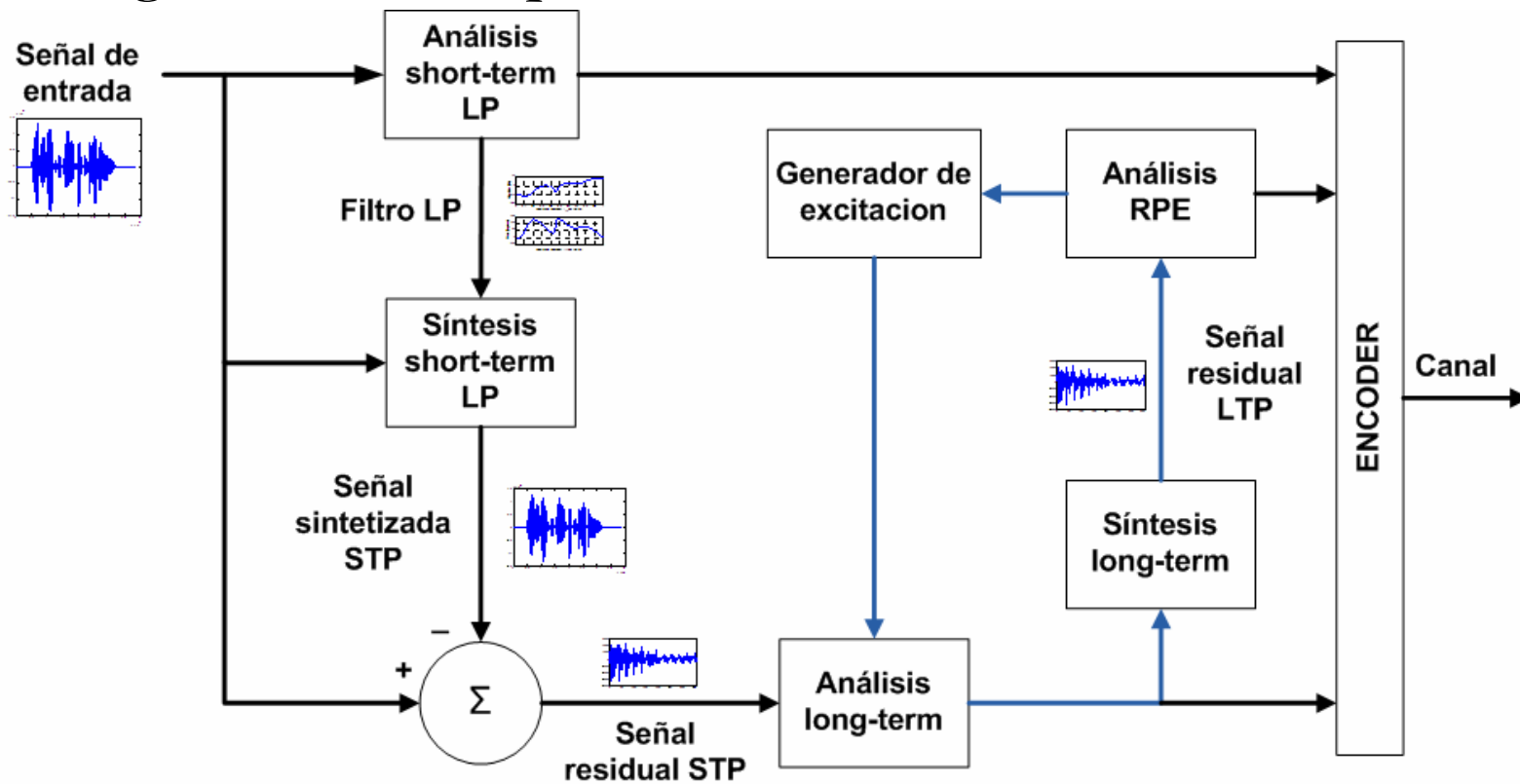
RPE-LPT (GSM)

□ Regular Pulse Excitation



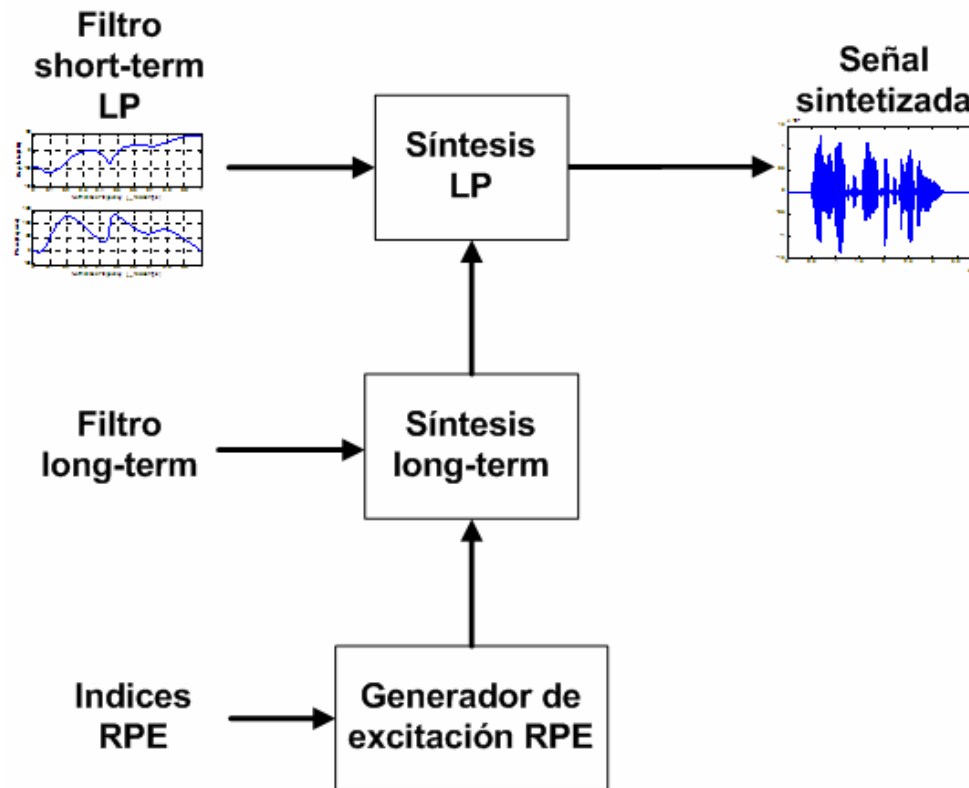
RPE-LPT (GSM)

□ Diagrama de Bloques - Encoder



RPE-LPT (GSM)

□ Diagrama de Bloques - Decoder

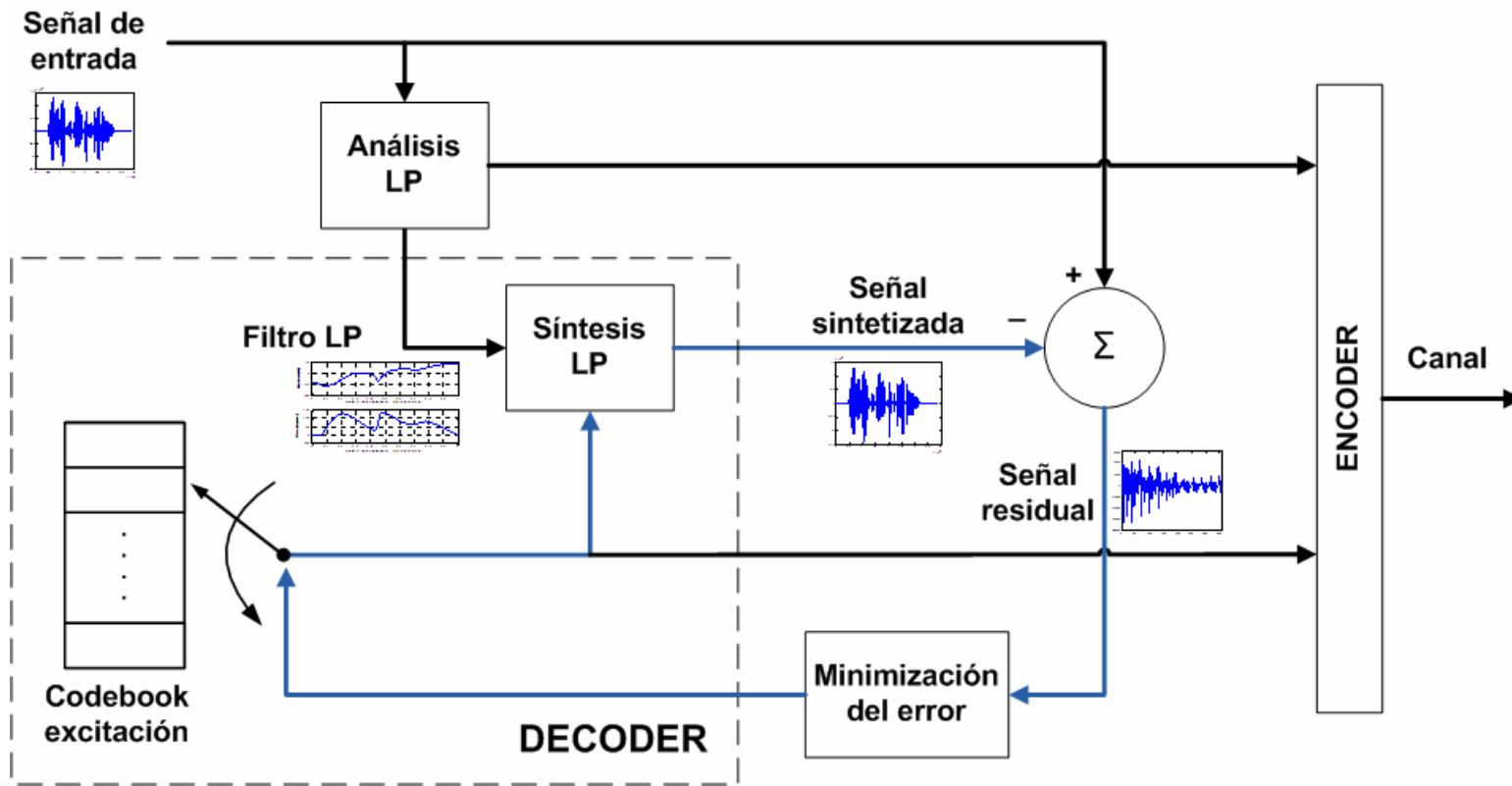


CELP

- Code-Excited Linear Predictive Vocoder
 - Short term predictor: se realiza análisis LPC-10
 - Long term predictor: se realiza análisis de periodicidad.
 - Se utiliza un **codebook** de señales de excitación para usar como corrección en la estimación de periodicidad.
 - La excitación utilizada es la que minimiza la energía de la señal residual.
 - Búsqueda exhaustiva.

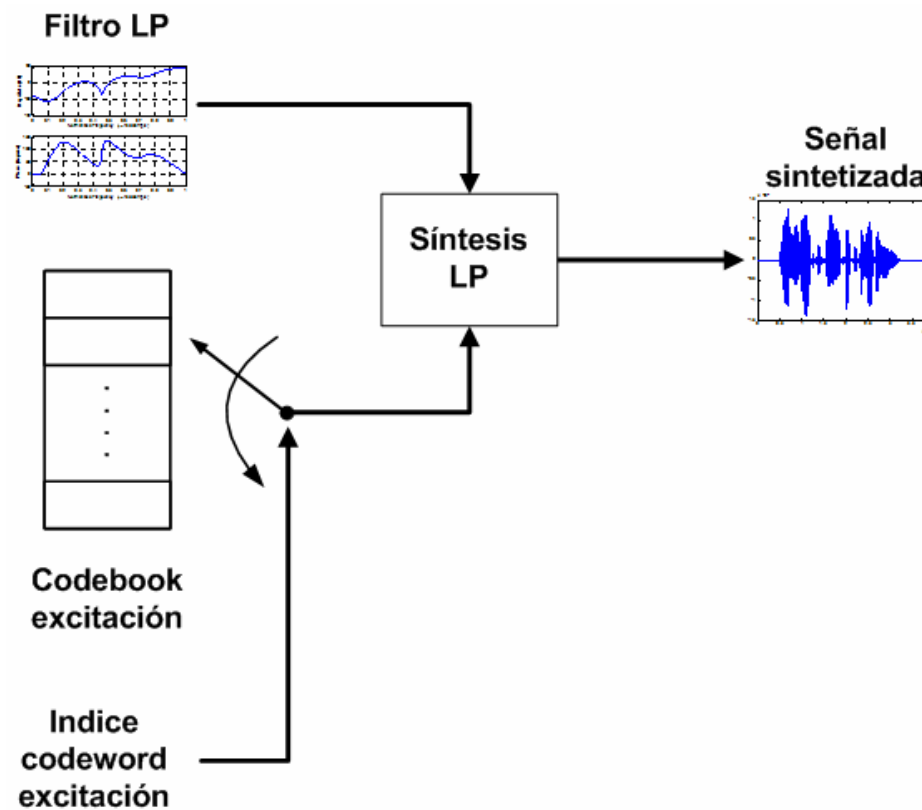
CELP

□ Diagrama de Bloques - Encoder



CELP

□ Diagrama de Bloques - Decoder



Análisis por Síntesis

□ Tabla comparativa

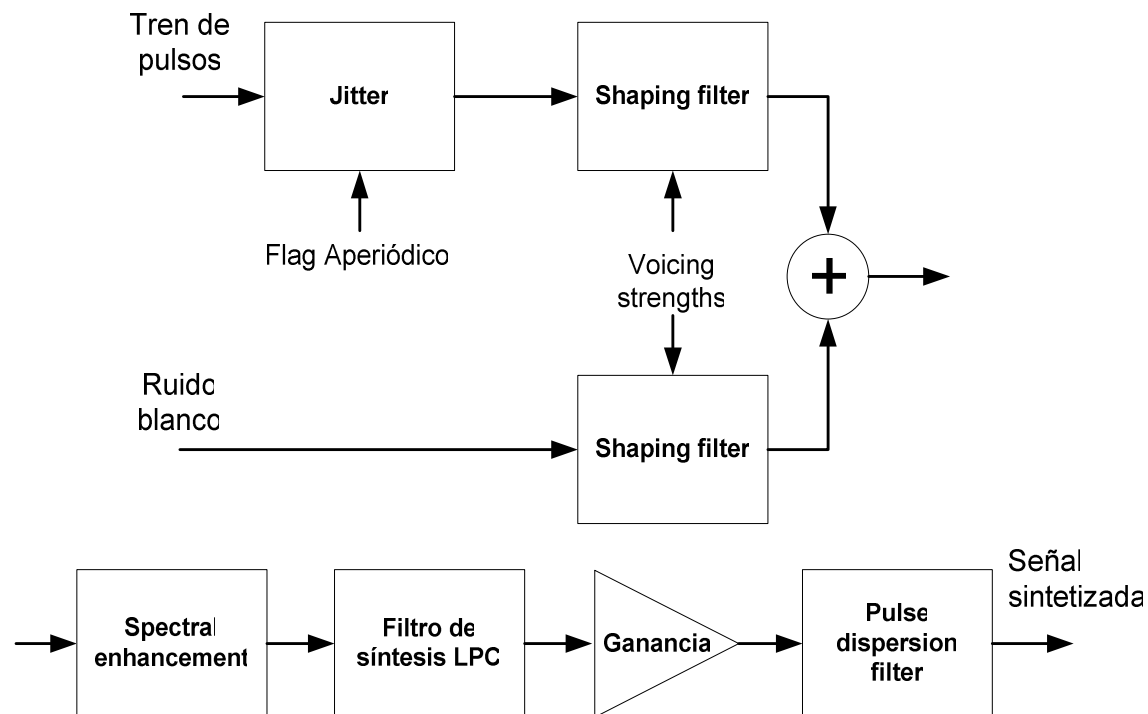
Técnica	Tasa de transmisión	MOS
PCM (G.711)	64 Kbps	4.4
ADPCM (G.726)	16 Kbps	4.1
LPC-10 (USFS 1015)	2.4 Kbps	2.3
RPE-LTP (GSM 6.10)	13 Kbps	3.5
CELP	4.8 Kbps	3.2

MELP

- Mixed-excitation linear prediction vocoder
- Está basado en el vocoder LPC tradicional, incluyendo cuatro mejoras
 - ▣ Mixed excitation
 - ▣ Pulsos aperiódicos
 - ▣ Spectral enhancement
 - ▣ Pulse dispersion filter

MELP

□ Diagrama del modelo de producción de voz

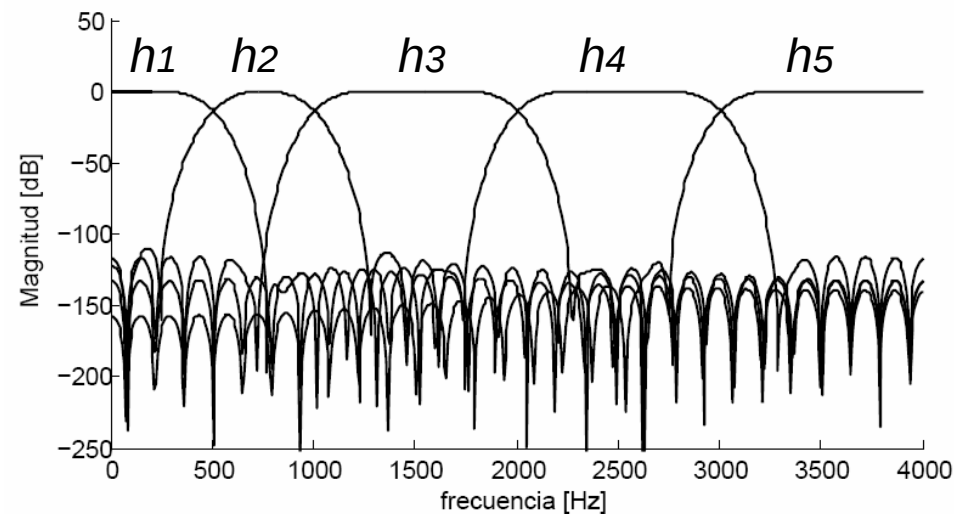


MELP

- Mixed excitation
 - ▣ Un gran problema vocoder LPC convencional es el cambio abrupto entre excitación y ruido blanco
 - ▣ MELP propone una mezcla entre excitación periódica y ruido
 - ▣ La cantidad de excitación periódica y ruido en la señal sintetizada en la mezcla depende de dos shaping filters

MELP

- Mixed excitation
 - ▣ Los shaping filters se construyen a partir de combinaciones lineales de filtros FIR pasabanda fijos llamados “Filtros de síntesis”

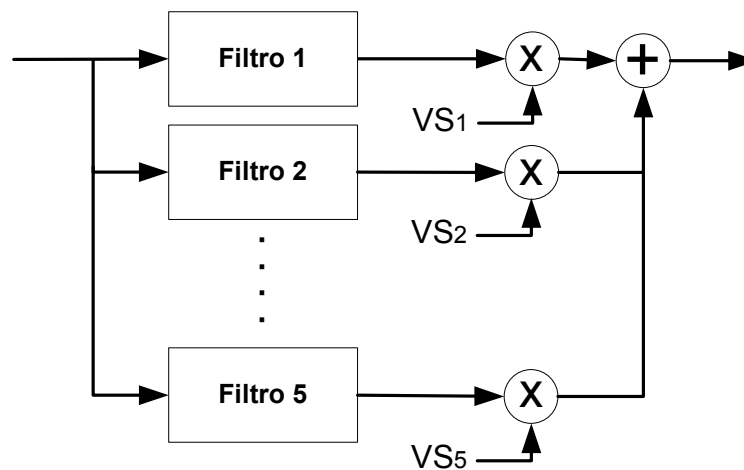


MELP

- Mixed excitation

- ▣ Si $h_i(n)$ es la respuesta al impulso del filtro de síntesis i , El shaping filter para excitación está definido como:

$$h_{pulso}(n) = \sum_{i=1}^5 v s_i h_i(n) \quad v s_i \in [0,1]$$



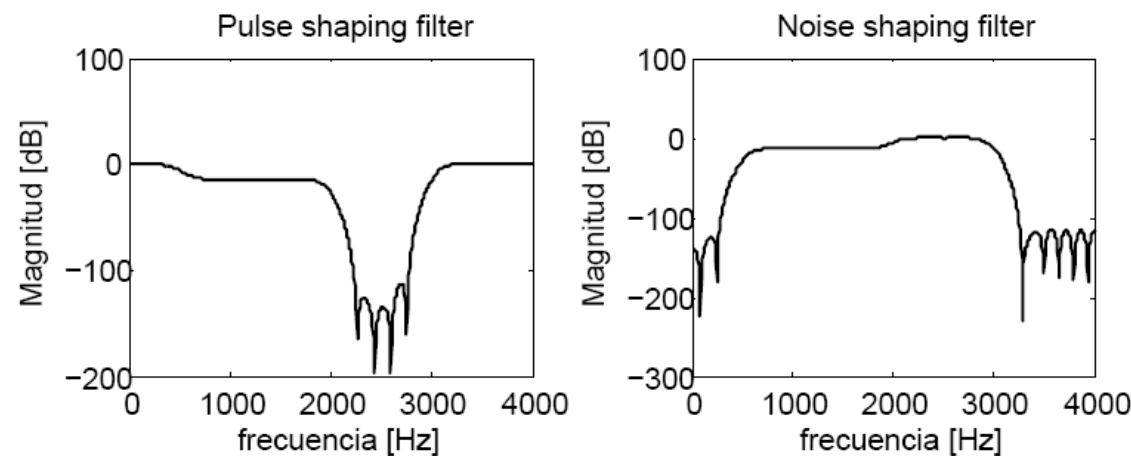
MELP

- Mixed excitation

- ▣ Análogamente, el shaping filter de ruido blanco está dado por:

$$h_{noise}(n) = \sum_{i=1}^5 (1 - vs_i) h_i(n)$$

- ▣ Ejemplo: Si $vs_1=1$, $vs_2=vs_3=0.5$, $vs_4=0$, $vs_5=1$ entonces:

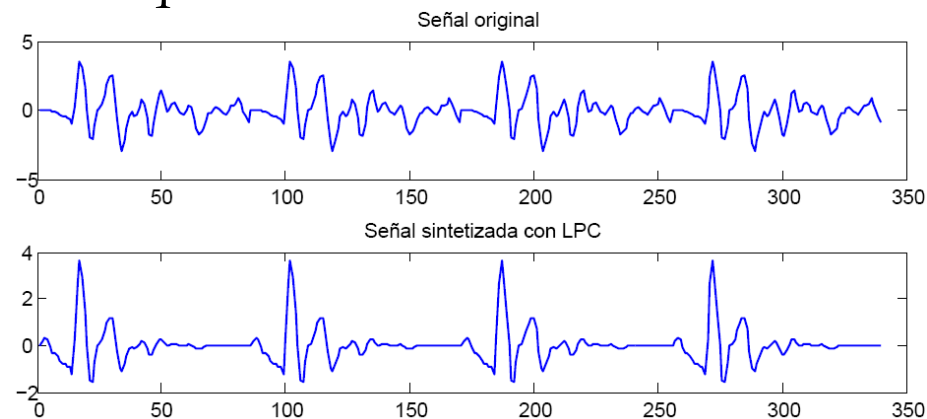


MELP

- Pulsos aperiódicos
 - ▣ Modela las transiciones de pitch y los fenómenos como el vocal fry (o creaky voice)
 - ▣ La idea es generar pulsos periódicos con jitter cuando sea necesario
 - ▣ Un frame es detectado como jittery voiced si $vs_1 < 0.5$
 - ▣ De esta forma, la decisión V/UV es menos exigente y se transforma en V/JV/UV
 - ▣ Se necesita sólo un bit para saber si un frame posee jitter o no

MELP

- Adaptive spectral enhancement
 - ▣ Filtro que mejora la calidad de la señal sintetizada
 - Los pulsos generados a partir de una señal sintetizada con LPC presenta un decaimiento más abrupto que una señal original
 - ▣ Se ajusta solamente de acuerdo a los coeficientes LPC, por lo cual no requiere bits extra



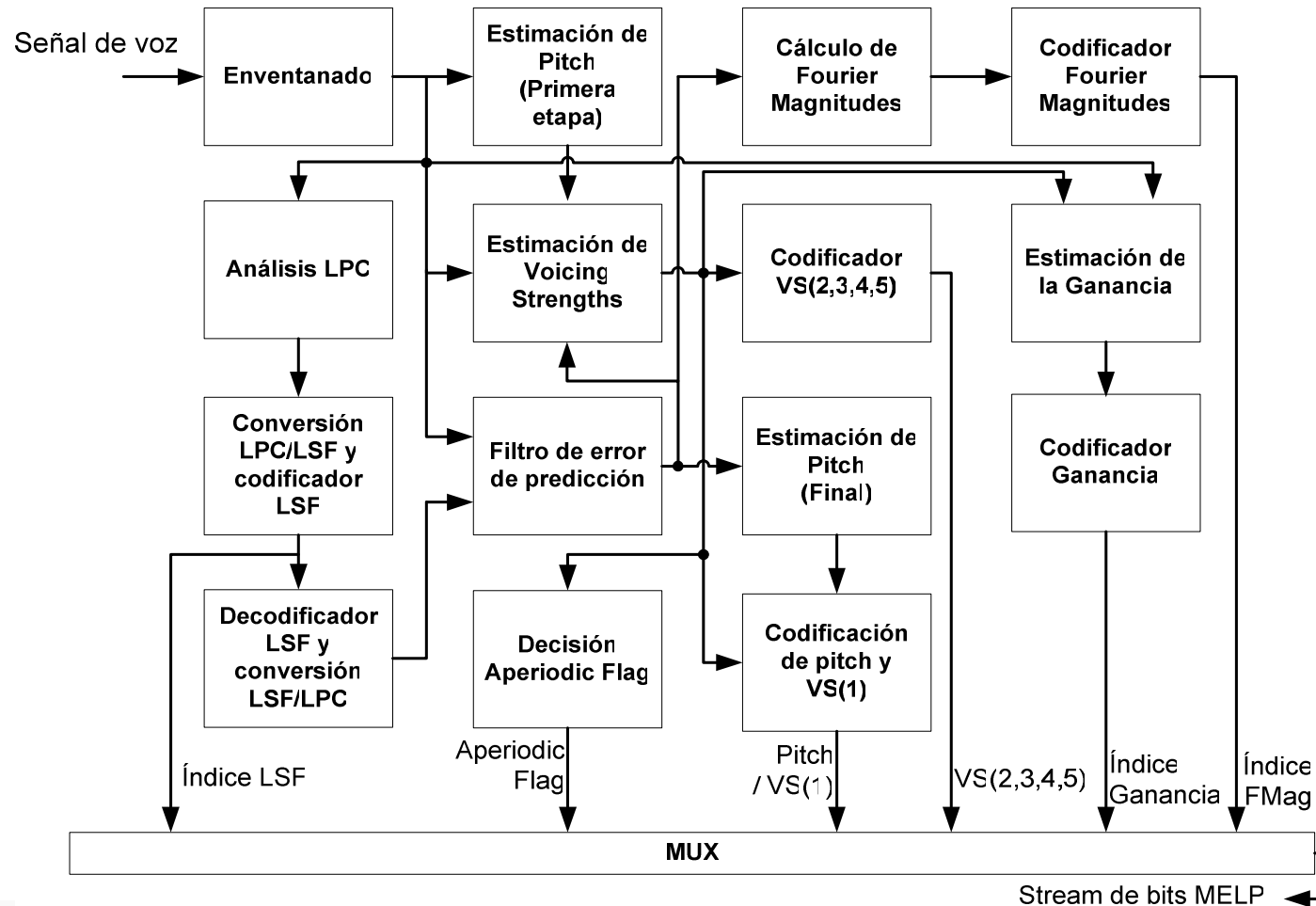
MELP

- Pulse dispersion filter
 - ▣ Mejora el matching entre la señal original y su versión sintetizada original en frecuencia
 - ▣ Consiste en un filtro FIR fijo
 - ▣ Al igual que spectral enhancement, no requiere gasto de bits extra

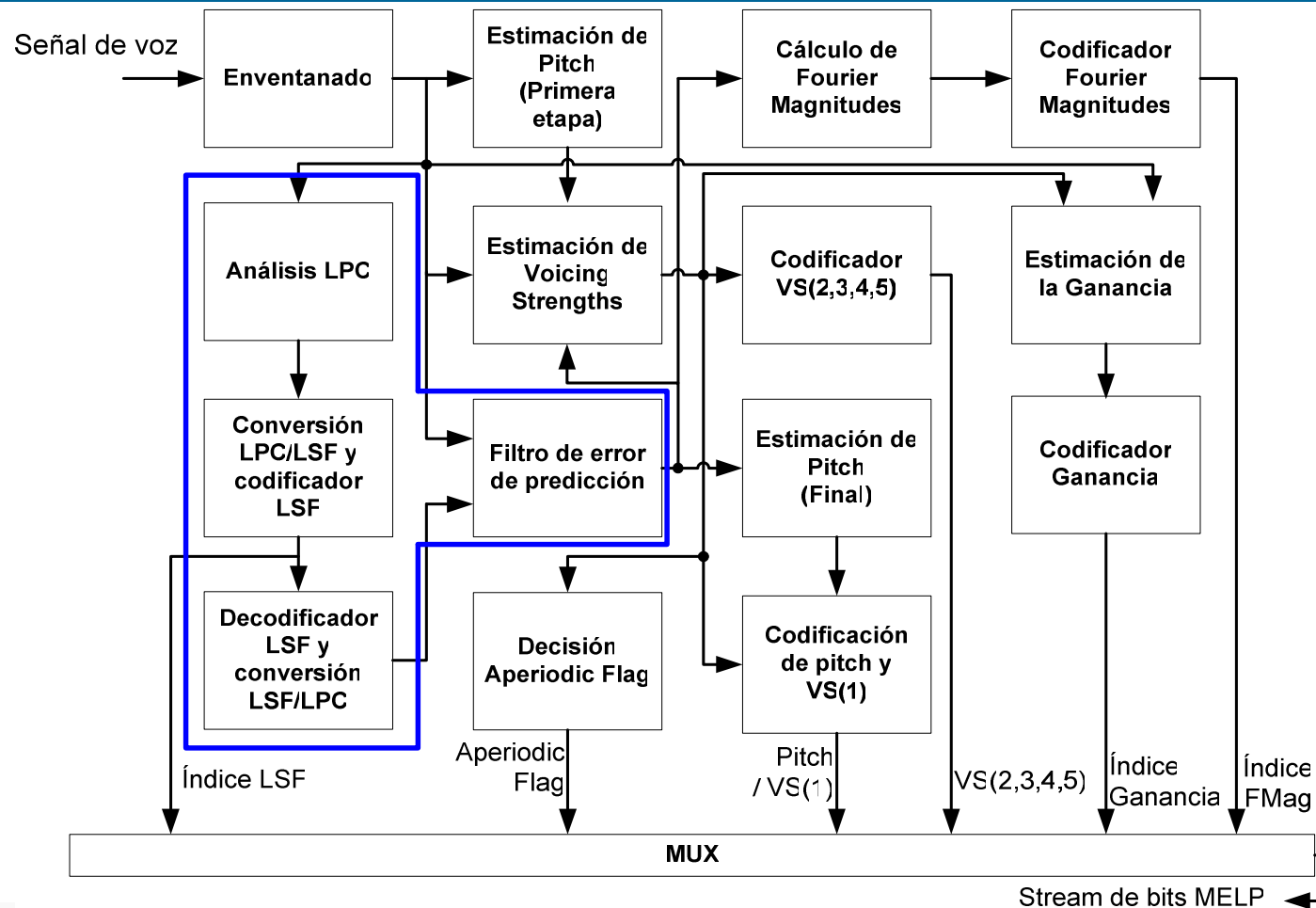
MELP

- Federal Standard MIL-STD-3005
 - ▣ Funcionamiento a bajas tasas de transmisión (2.4 kbps)
 - ▣ Robusto en ambientes adversos (ruido)
 - ▣ Eficiente en términos de complejidad computacional

Codificador MELP

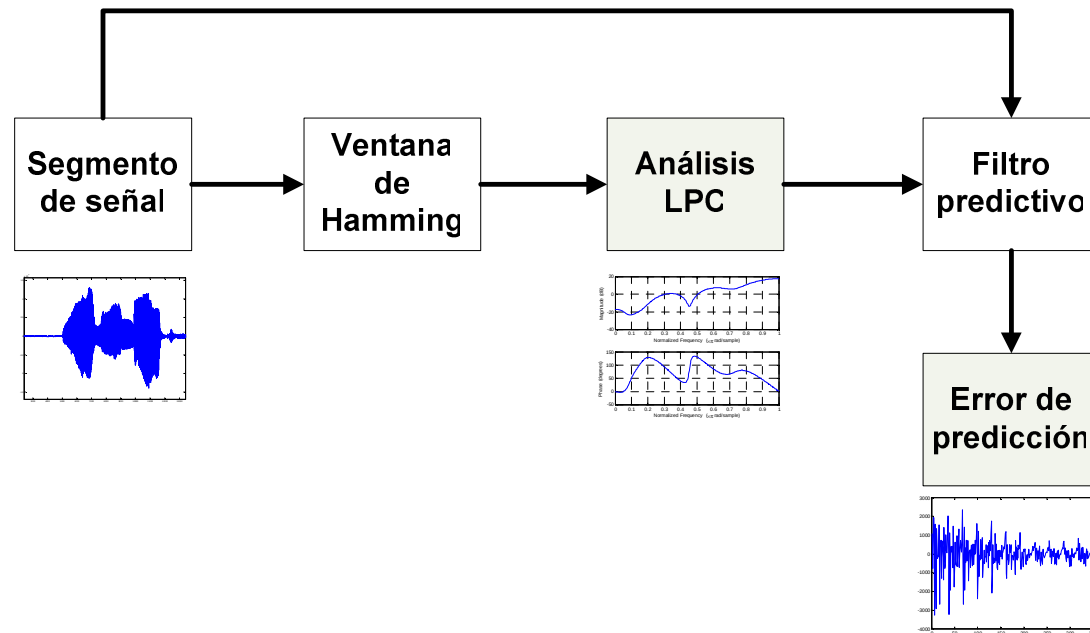


Codificador MELP

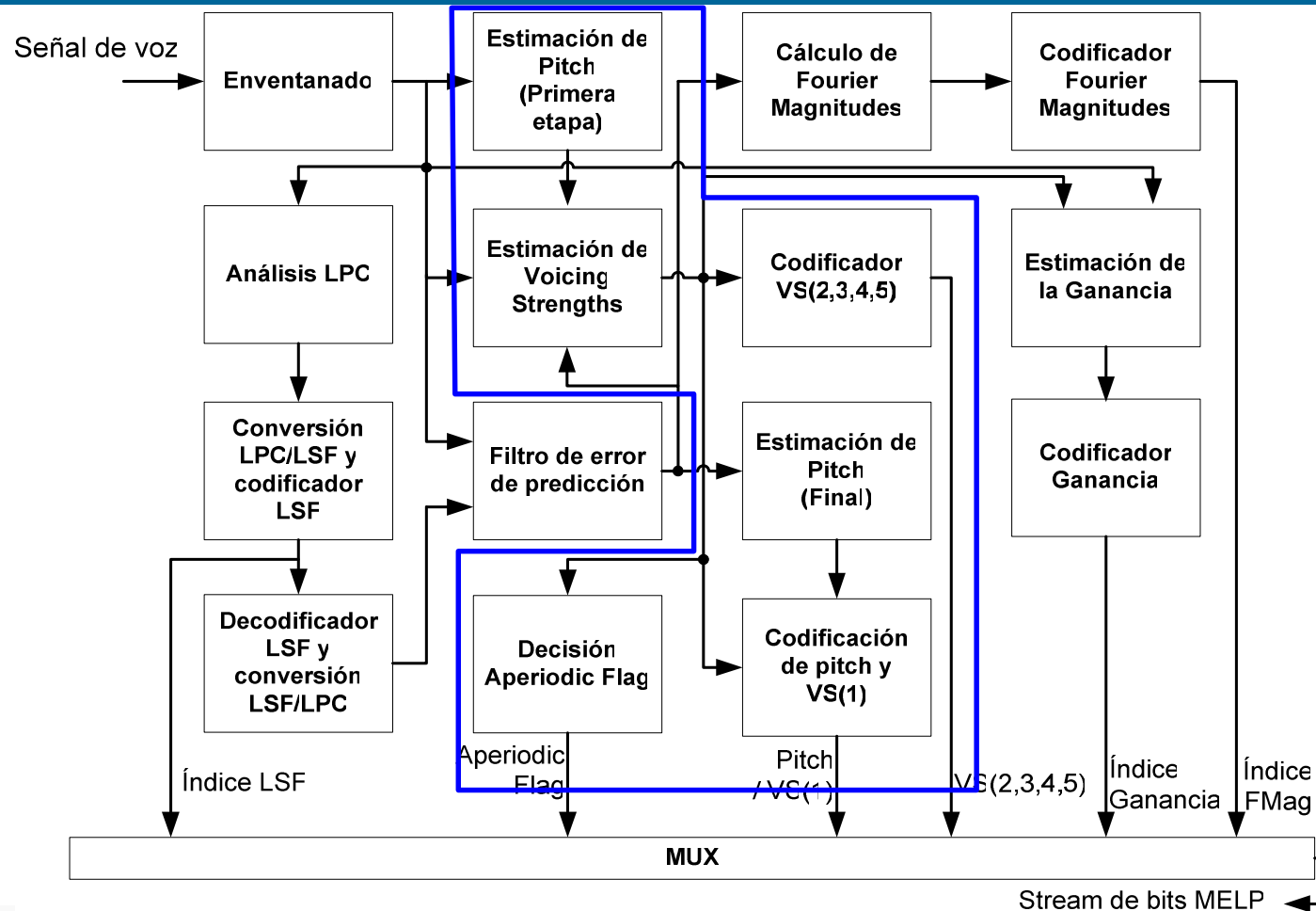


Contenidos

□ Bloque de análisis LPC



Codificador MELP



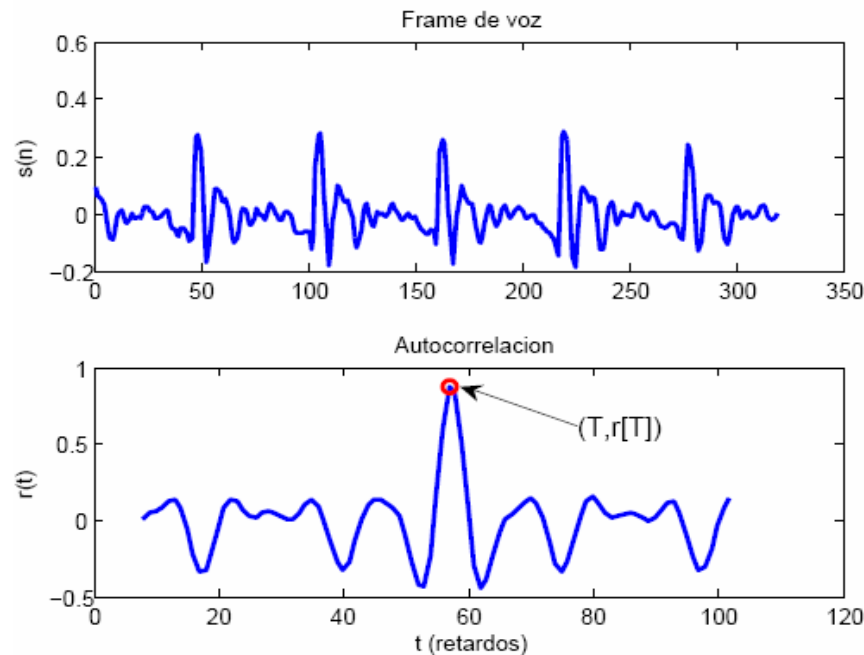
Codificador MELP

- Voicing strengths analysis y estimación de pitch
 - ▣ Los valores de V_{Si} resultan del análisis de peridiocidad de cada una de las bandas dadas por los filtros de síntesis
 - ▣ La peridiocidad se mide a través de la función de autocorrelación:

$$r(t) = \frac{\sum_{n=0}^{N-1} s(n)s(n+t)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}}$$

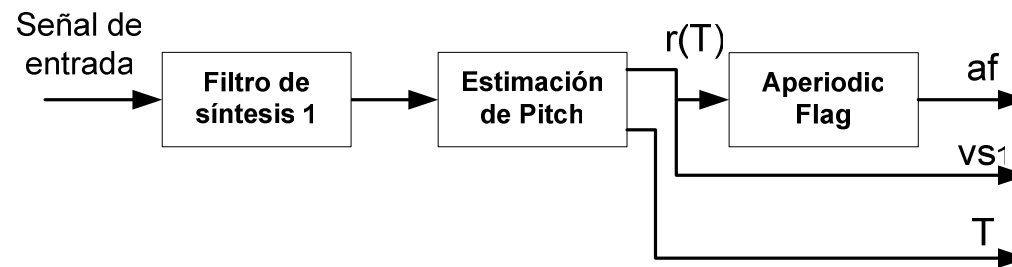
Codificador MELP

- Voicing strengths analysis y estimación de pitch
 - ▣ El retardo T que maximiza $r(t)$ corresponde al pitch. Además, el valor de la $r(T)$ representa el grado de periodicidad de $s(n)$



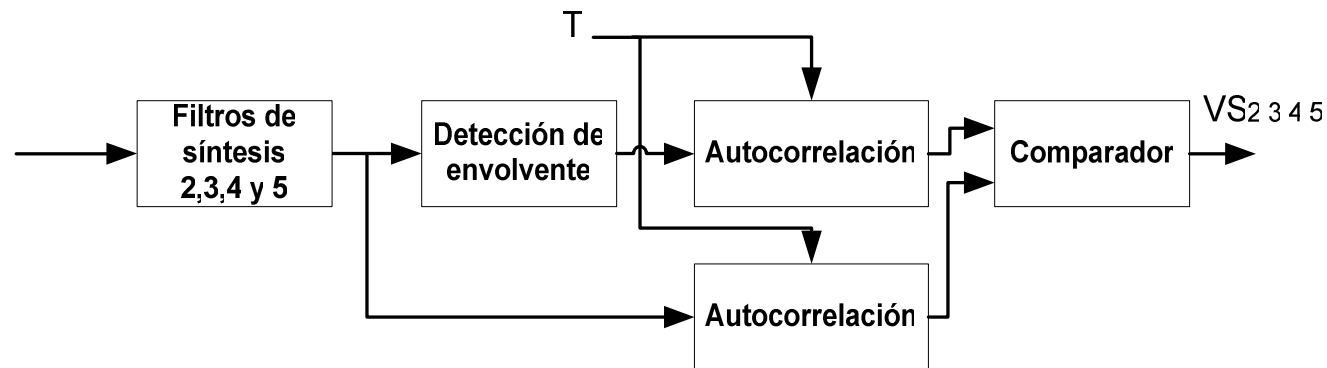
Codificador MELP

- Voicing strengths
 - ▣ VS_1 se calcula de acuerdo a:

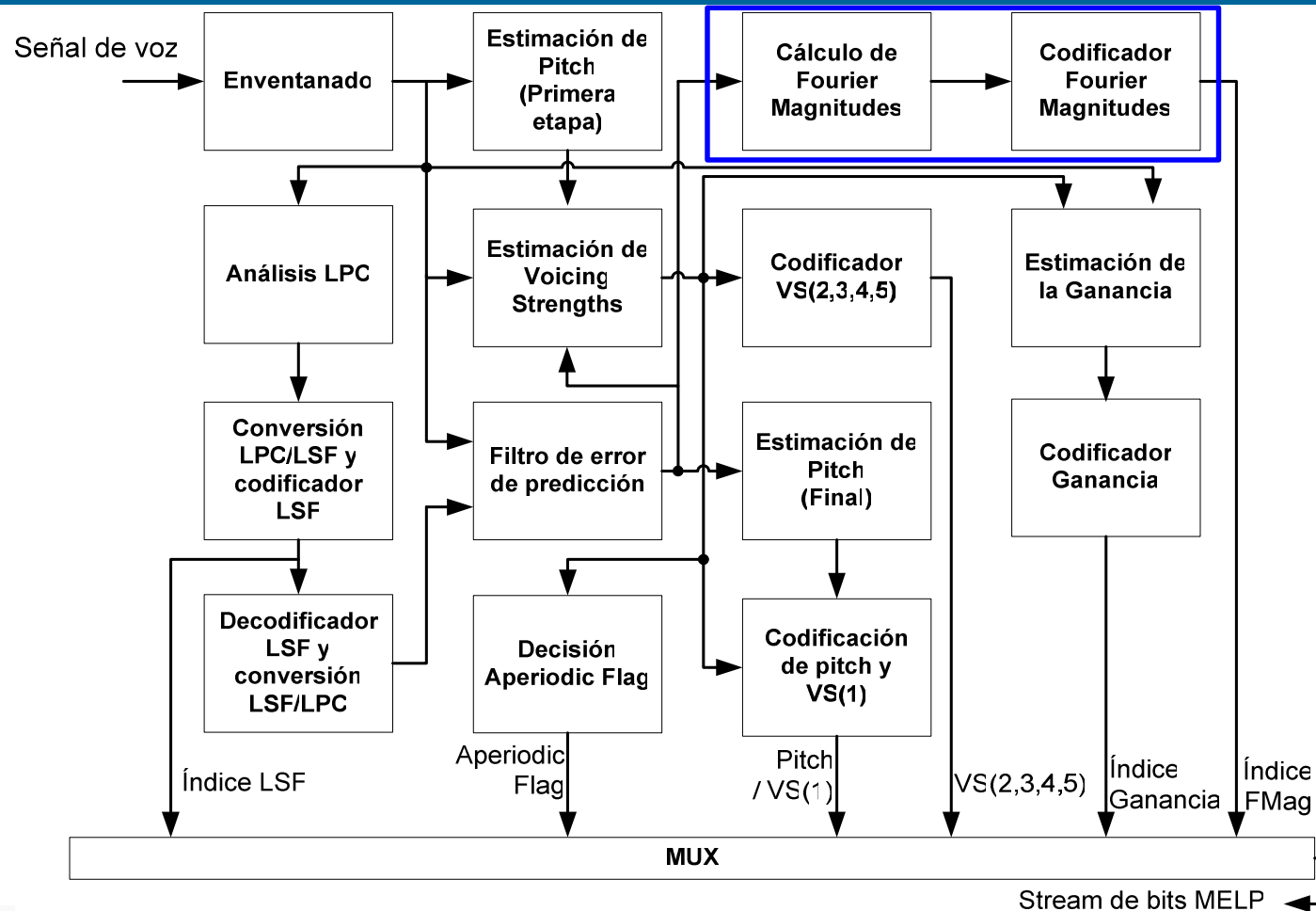


MELP

- Voicing strengths
 - ▣ Los restantes voicing strengths, $VS_{(2,3,4,5)}$, se estiman según:



Codificador MELP

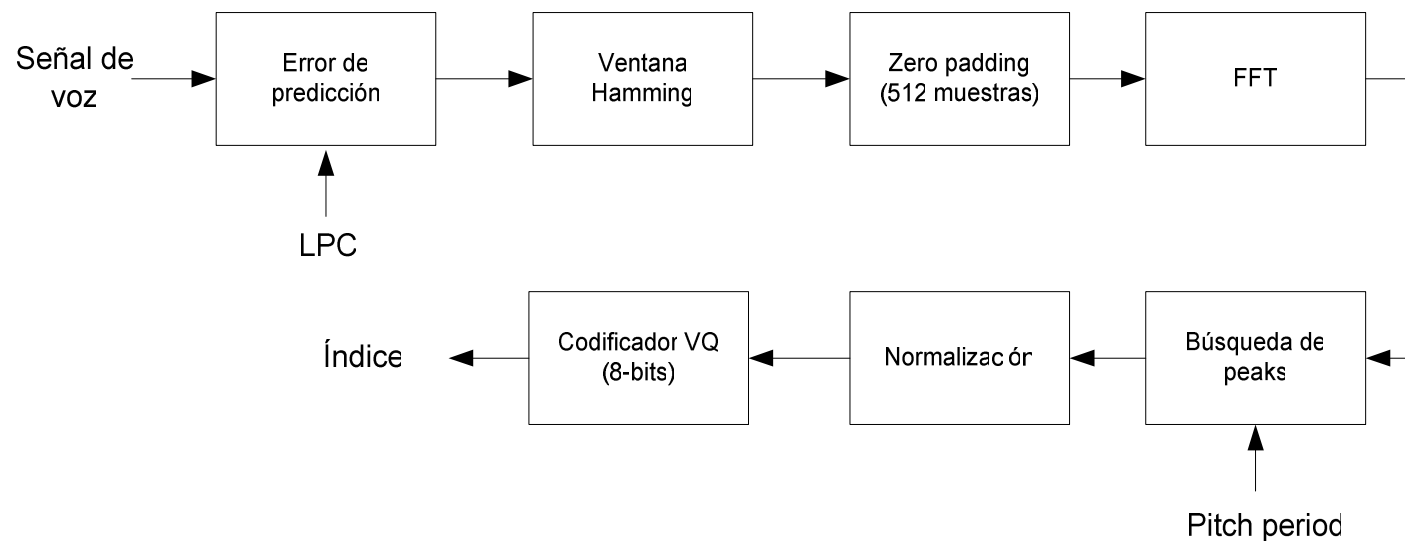


Codificador MELP

- Fourier magnitudes
 - ▣ Es la magnitud de la transformada de Fourier del error de predicción.
 - ▣ Se captura la forma del pulso de excitación
 - ▣ El objetivo es agregar en la excitación de la señal sintetizada información del residuo del error de predicción
 - Análisis por síntesis

Codificador MELP

- Cálculo Fourier magnitudes
 - ▣ Diagrama general



Codificador MELP

- Calculo Fourier magnitudes
 - Entradas:
 - Error de predicción
 - Pitch
 - Enventanado Hamming al error de predicción
 - Aplicación Zero padding a 512
 - Estimación de FFT sobre las 512 muestras.

Codificador MELP

- Calculo Fourier magnitudes
 - Búsqueda de armónicos mediante peak de magnitud en la FFT
 - Se buscan peak en las vecindades de los 10 primeros armónicos del pitch
 - Normalización de los Fourier magnitudes
 - La secuencia de Fourier magnitudes normalizada tiene un valor RMS=1
 - La normalización se realiza como:

$$F'_{mag}[i] = \alpha \cdot F_{mag}[i]$$

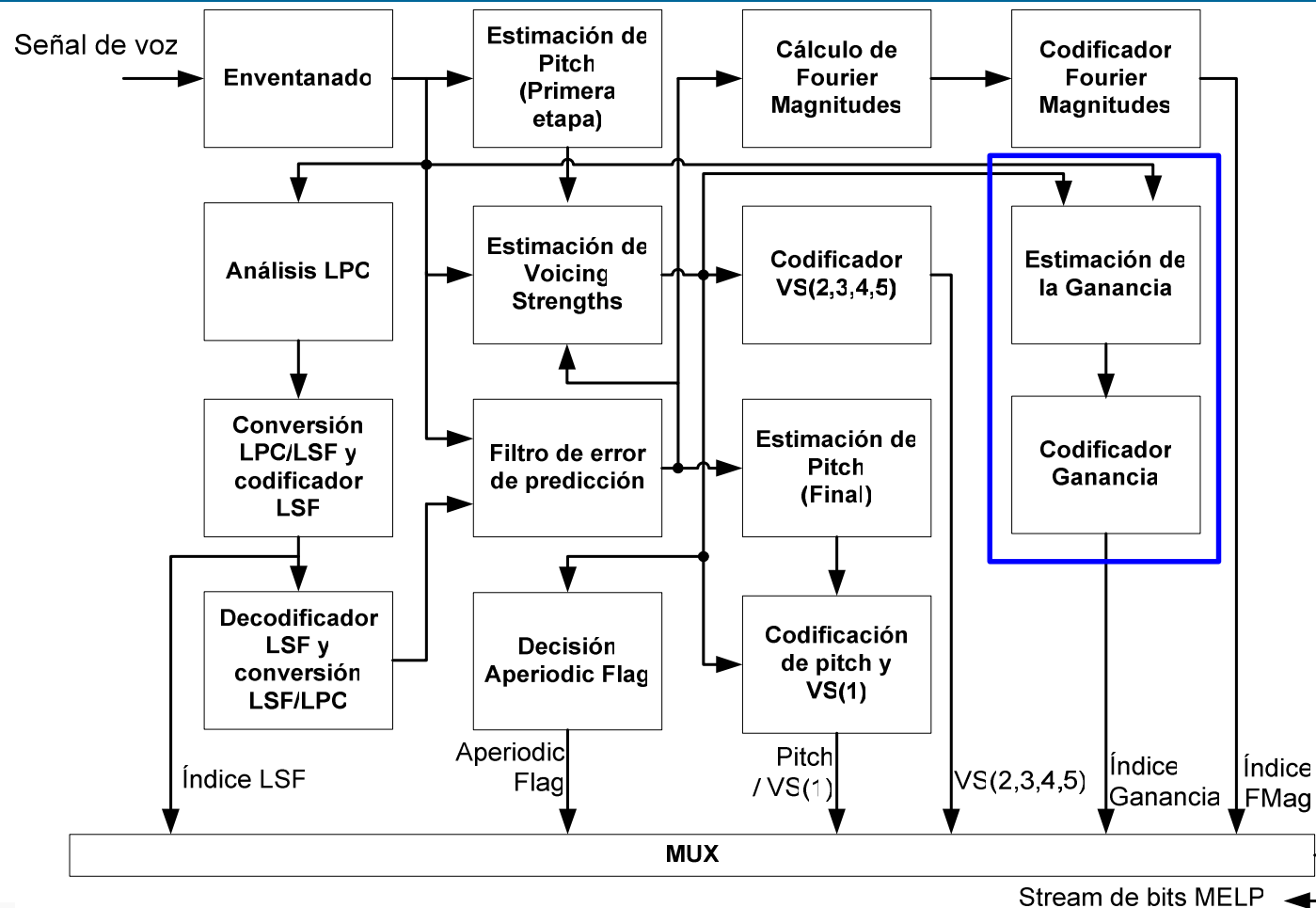
$$\alpha = \left(\frac{1}{10} \sum_i (F_{mag}[i])^2 \right)^{\frac{1}{2}}$$

Codificador MELP

- Calculo Fourier magnitudes
 - ▣ Cuantización de los Fourier magnitudes
 - Se utiliza una distancia euclidiana ponderada y un codebook de 8 bits
 - Los pesos de cada Fourier magnitudes es estima como:

$$w_i = \left[\frac{117}{23 + 75 \left(1 + 1.4 \left(\frac{F_{mag}[i]}{1000} \right)^2 \right)^{0.69}} \right] \quad \begin{array}{l} i = 1 \dots 10 \\ F_{mag}[i] = \text{Fourier Magnitude } i \end{array}$$

Codificador MELP



Codificador MELP

□ Ganancia

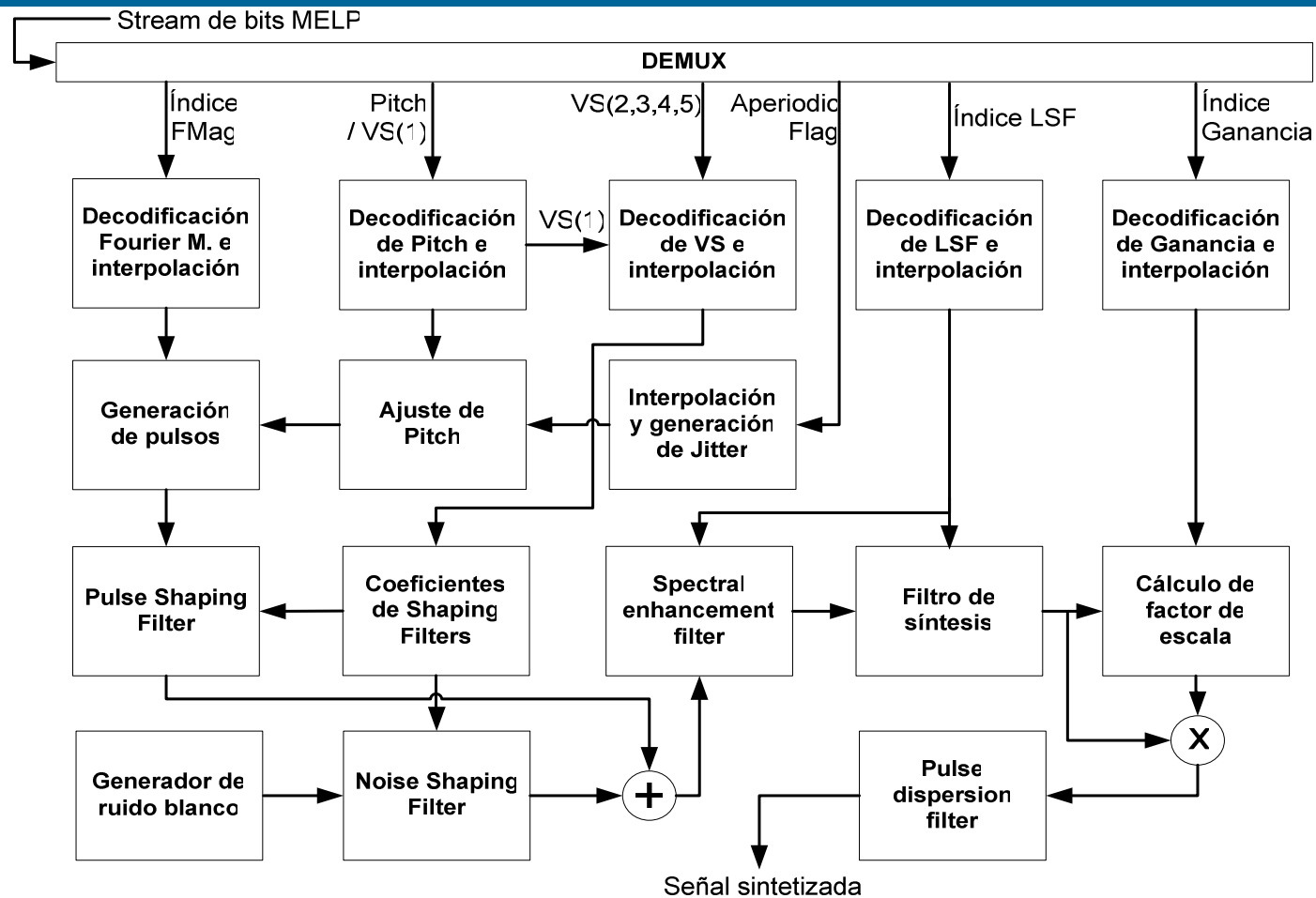
- Se calculan dos ganancias por frame (G_1 y G_2)
- Se utilizan ventanas variables para la señal de entrada que dependen del pitch
- Estimación de las ganancias:

$$G_i = 10 \log_{10} \left(0.01 + \frac{1}{L} \sum_{n=1}^L s_n^2 \right) \quad L = \text{Tamaño de la ventana}$$

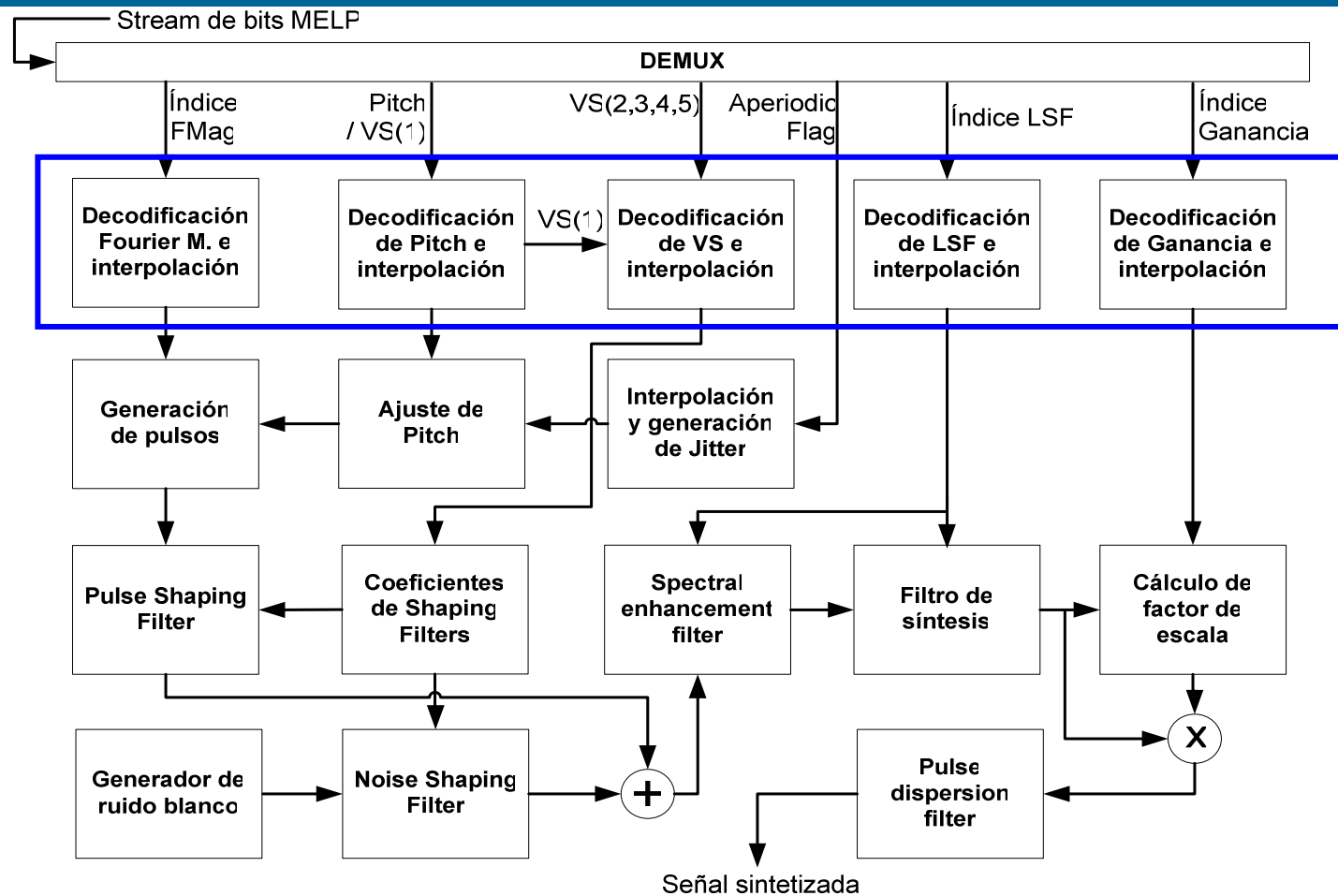
Codificador MELP

- Ganancia
 - ▣ Cuantización de la Ganancia
 - Cuantización lineal para G_1 y G_2
 - G_1 se cuantiza con 3 bits
 - G_2 se cuantiza con 5 bits

Decodificador MELP



Decodificador MELP



Decodificador MELP

□ Interpolación

■ Se interpolan todos los parámetros de síntesis

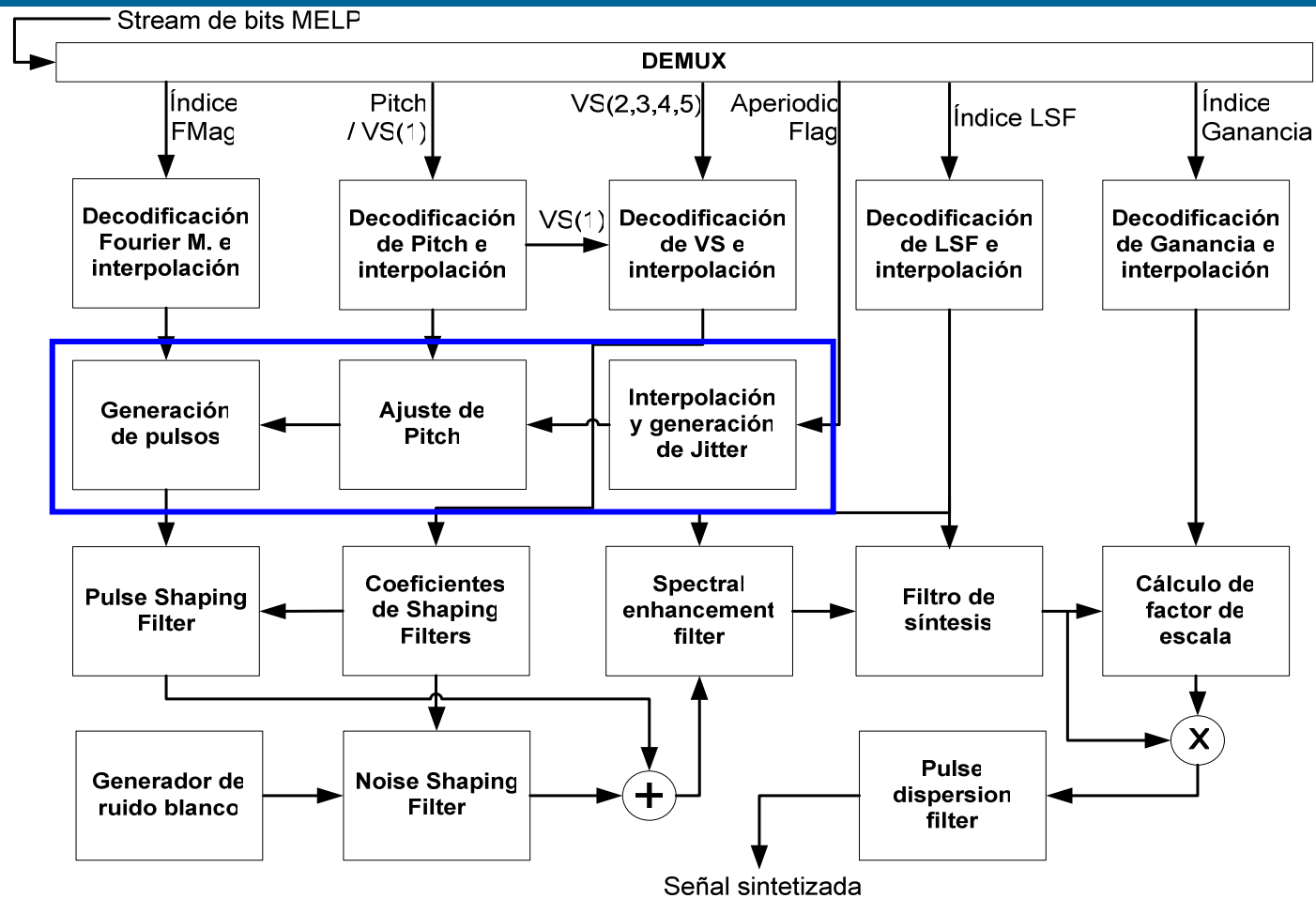
- Pitch, LSF, jitter, Fourier magnitudes y voicing strengths son interpolado de la siguiente manera:

$$T = (1 - \alpha)T_{past} + \alpha T_{present}$$

- La ganancia se interpola de manera diferente:

$$g = \begin{cases} (1 - \alpha)g_{2_{past}} + \alpha g_{1_{present}} & n_0 < 90 \\ (1 - \alpha)g_{1_{present}} + \alpha g_{1_{present}} & 90 \leq n_0 < 180 \end{cases}$$

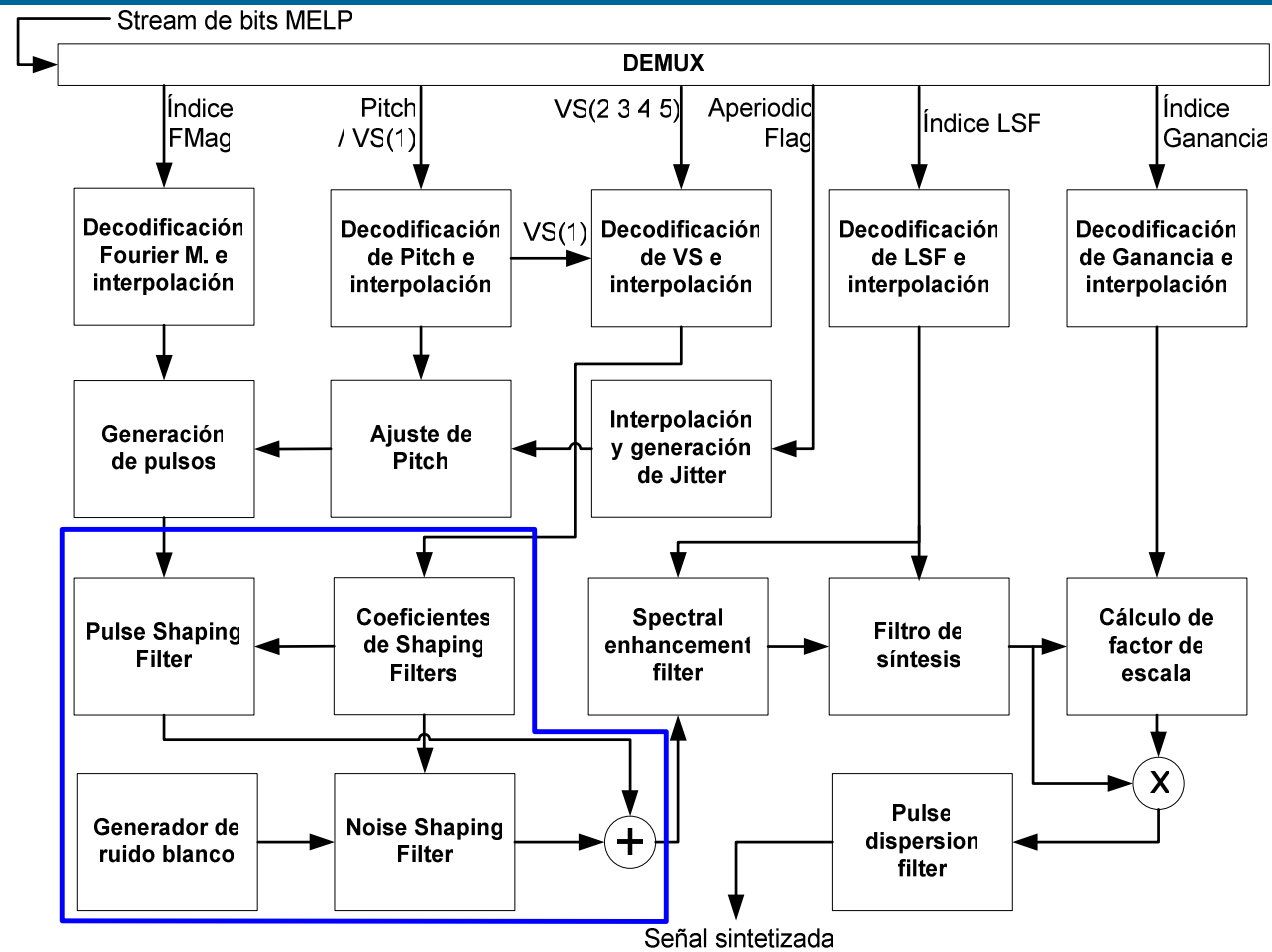
Decodificador MELP



Decodificador MELP

- Generación de pulsos
 - ▣ La señal sintetizada se genera a partir de pulsos
 - Su forma depende de los Fourier Magnitudes
 - Su largo (período) depende del valor del pitch y el jitter
 - Si el no existe pitch (caso Unvoiced) se utilizan valores por defecto

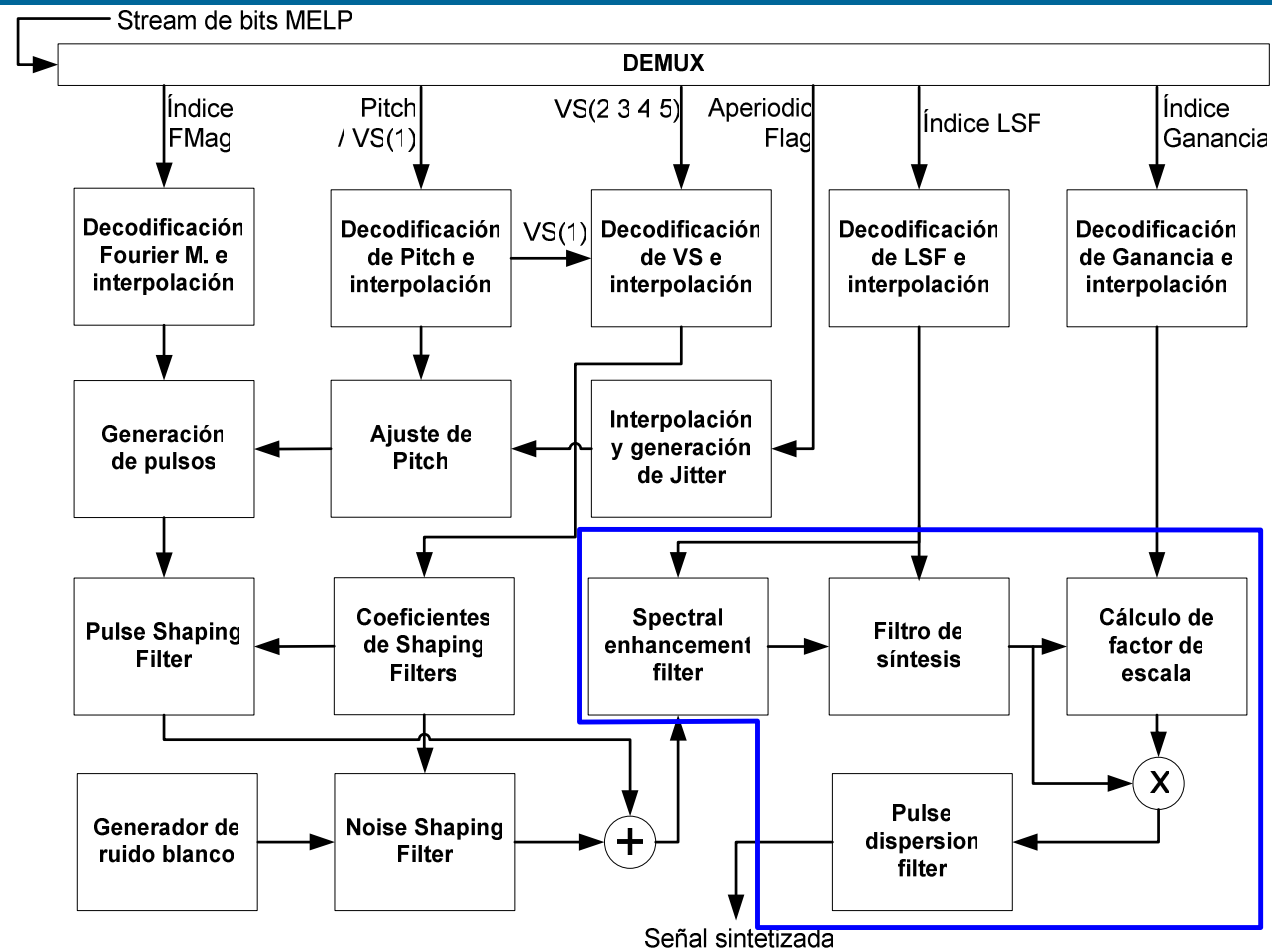
Decodificador MELP



Decodificador MELP

- Generación de excitación mezclada
 - Los voicing strengths interpolados toman valores en el intervalo $[0,1]$
 - Ambos shaping filters se generan a partir de sumas ponderadas de filtros de síntesis
 - VSi para el shaping filter de excitación
 - $(1-VSi)$ para el shaping filter de ruido
 - Luego se filtra la excitación (pulsos) y el ruido
 - Finalmente, se suma la excitación y el ruido ya filtrados

Decodificador MELP



MELP

□ Bit allocation

Parámetro	Voiced	Unvoiced
LSF	25	25
Pitch y VS1	7	7
VS(2,3,4,5)	4	-
Primera ganancia	3	3
Segunda ganancia	5	5
Aperiodic flag	1	-
Fourier magnitudes	8	8
Sincronización	1	1
Protección de errores	-	13
TOTAL	54	54

MELP

□ Tabla comparativa

Técnica	Tasa de transmisión	MOS
PCM (G.711)	64 Kbps	4.4
ADPCM (G.726)	16 Kbps	4.1
LPC-10 (USFS 1015)	2.4 Kbps	2.3
RPE-LTP (GSM 6.10)	13 Kbps	3.5
CELP	4.8 Kbps	3.2
MELP	2.4 Kbps	2.9