UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

FACULTAD DE INGENIERÍA

DISEÑO DE CODIFICACIÓN DE VOZ PARA TELEFONÍA CELULAR GSM

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN INGENIERÍA

ELECTRICA - TELECOMUNICACIONES

P R E S E N T A :

JUDITH VALDES CORDERO

TUTOR:
DR. JOSÉ ABEL HERRERA CAMACHO

2006

Autorizo a la Dirección General de Bibliotecas de la UNAM a difundir en formato electrónico e impreso el contenido de mi trabajo recepcional.
NOMBRE: Judith Valdés Cardero
FECHA: 31- mayo - 2006
FIRMA:

JURADO ASIGNADO:

Presidente: Dr. Bohumil Psenicka

Secretario: Dr. Luis Alberto Pineda Cortes

Vocal: Dr. Abel Herrera Camacho

1er. Suplente: Dr. Javier Gómez Castellanos

2do. Suplente: Dr. Francisco García Ugalde

Lugar o lugares donde se realizó la tesis:

Facultad de Ingeniería, UNAM

TUTOR DE TESIS:
Dr. José Abel Herrera Camacho

FIRMA

A mi Padre, quien no vivió lo suficiente para ver este trabajo.

AGRADECIMIENTOS.

A la Universidad Nacional Autónoma de México y en especial a la Facultad de Ingeniería.

A mi madre, quien no solo es mi sustento, sino que se ha convertido en una verdadera amiga.

A mis hermanos Ernesto y Rosa María que además de ser mi mejor ejemplo ha seguir, son mis mas gratos momentos vividos. A Ale, Ray A.A. y Ray A.V. por ser parte se esos momentos.

A mi tía Judith Cordero, por su continuo apoyo en cada cosa que hago, por las comidas y el cine los sábados, a Judy Campos por ser mi tercera hermana.

A mi maestro Abel por darme la oportunidad de hacer lo que alguna vez soñé y por supuesto por apoyar y ser parte de este trabajo.

A mis amigos, Diego, Enrique, Francisco, Goyo, Jorge, José Luis, Rita y Ximena, porque aunque la distancia nos separe, siempre serán un motivo para hacer las cosas, por que los quiero.

A Gilberto por el profundo amor que le tengo y por creer juntos que siempre podemos ser mejores.

A Carolina Galindo por creer en mi, por las platicas sinceras y por el apoyo otorgado para llevar a cabo el término de este trabajo

Al CONACYT, por su labor, por el apoyo otorgado en mis estudios de maestría que permitieron la elaboración de este trabajo.

ÍNDICE.

Objetivos

Como objetivos, se marcan:

- Analizar los diferentes esquemas de codificación de voz utilizadas en GMS.
- Con base en los métodos aplicados en diversos estándares, poder diseñar un codificador que funcione para la telefonía celular GSM, mejorando la calidad de voz actual.
- Simular tanto el estándar de la telefonía celular GSM como los algoritmos diseñados.
- Realizar pruebas para poder determinar la calidad de voz de los algoritmos diseñados y determinar si se mejoro la calidad del codificador actual usado en la telefonía GSM.

RESUMEN

La necesidad del ser humano de comunicarse ha llevado al desarrollo de grandes tecnologías en el ámbito de las telecomunicaciones y en especial en la telefonía celular donde el procesamiento de voz tiene una de sus grandes aplicaciones y dentro del procesamiento de señales es una de las áreas más excitantes para la investigación.

En especial, los algoritmos de codificación de voz han permitido un incremento en la capacidad de las comunicaciones, es decir, han hecho a las comunicaciones de voz y al almacenamiento de datos de voz más efectivos y eficientes.

Uno de los sistemas de telefonía celular más exitosos y con más demanda es el estándar GSM (*Global System for Mobile communications*) que por diseño puede aceptar diferentes algoritmos de codificación de voz, sin embargo, los sistemas de comunicación trabajan con una velocidad de transmisión fija que no puede ser alterada sin modificar drásticamente el diseño además de tener una alta calidad de voz.

Para GSM el estándar es RPE-LTP (*Regular Pulse Excitation-Long Term Prediction*), con el cual se obtiene una buena calidad de voz con una complejidad computacional no alta y una velocidad de transmisión de 13 kbps.

En este trabajo se desarrollaron cambios en el diseño del algoritmo de RPE-LTP para telefonía celular GSM respetando la estructura general para no alterar la velocidad de transmisión pero logrando mejoras en la calidad de voz.

De manera general la estructura de diseño es: Un pre-procesado de la señal aplicando diferentes filtros y ventanas, un análisis y filtro de STP (*Short Term Prediction*) aplicando de igual manera diferentes métodos para la obtención de los coeficientes del filtro como: Levinson Durbin, Recursion de Schur y algunas modificaciones del primero, esto para eliminar la correlación entre tramas, posteriormente la calidad es mejorada con una etapa de análisis y filtro de LTP (*Long Term Prediction*) y por último aplicando una estructura periódica de la excitación para sonidos sonoros donde los principales cambios en el diseño se encuentran en la aplicación de diferentes métodos para la obtención de la frecuencia fundamental.

Todos los diseños hechos fueron simulados e implementados con ayuda del paquete de computo MATLAB y fueron probados junto con el estándar original con la medida subjetiva adecuada para

este tipo de codificadores (codificadores híbridos) llamada MOS (*Mean Opinion Score*) determinando la calidad de los algoritmos de codificación de voz y obteniendo para el estándar original un valor MOS de 3.7 mientras que los algoritmos modificados el valor MOS alcanza 3.9 y 4 por lo que se concluye que hubo mejoras en la calidad de voz.

CAPÍTULO 1.
ANTECEDENTES

La voz es la forma más natural que tenemos los seres humanos para comunicarnos y hoy más que nunca con la gran movilidad de las personas en el mundo, se han desarrollado grandes tecnologías en el ámbito de las telecomunicaciones y en especial en la telefonía celular, donde el procesamiento de voz tiene una de sus más grandes aplicaciones y dentro del procesamiento de señales es una de las áreas más excitantes para la investigación.

En especial, los algoritmos de codificación de voz han tenido una gran relevancia en los sistemas de telefonía digital celular al permitir un incremento en la capacidad de las comunicaciones, es decir, han hecho a la transmisión de voz y al almacenamiento de datos de voz más efectivos y eficientes [1]. Por lo que en este capítulo se presentaran los conceptos básicos de la codificación de voz como: las diferentes técnicas de codificación, estándares, requerimientos para el diseño y los métodos usados para determinar la calidad de los algoritmos de codificación de voz.

1.1 Tipos y estrategias de codificación de voz

Cuando la telefonía sobrepaso las expectativas en el servicio, se creo la necesidad de incrementar la capacidad en las comunicaciones telefónicas, es decir, aumentar el número de canales en las redes de telefonía que en ese entonces eran analógicas, así la solución de este problema se le conoció como la compresión del ancho de banda. Sin embargo la mayoría de las redes de telefonía existentes tanto fijas como móviles son digitales y la compresión de ancho de banda se convirtió en codificación de voz, la cual podemos conceptuar como la representación de la señal de voz en forma binaria (bits) y cuyo objetivo primordial es logra la mayor eficiencia, obteniendo el menor número de bits para ser enviados en un canal y poder reconstruir dicha señal con las menores perdidas en la calidad.

Existen varios tipos de codificación por lo que se ha creado una clasificación general según sus características, las cuales se presentan a continuación [2].

1.1.1 Codificadores de forma de onda (Waveform coders).- La principal característica de estos codificadores se basa en reproducir en el decodificador la señal de voz original muestra a muestra partiendo de una muestra inicial, estos tipos de codificadores no toma en cuenta la naturaleza de la señal y pueden ser operados tanto en el dominio del tiempo como en el dominio de la frecuencia.

Surgen 1937 cuando Reeves desarrolló uno de los sistemas más importantes, el *PCM* (Modulación por Pulsos Codificados), planteando el problema de la codificación eficiente de la señal de voz para reducir al máximo la velocidad de transmisión y el número de bits necesarios para la codificación binaria de la señal de voz. A partir de *PCM* se desarrollaron *DPCM* y *ADPCM*, algunos propuestos como estándares por la *ITU (International Telecommunications Union)*.

Proporcionan una alta calidad de voz pero con velocidades entre 32 kbps y 64 kbps, lo cual los hace poco utilizados cuando se necesitan *bit rates* bajos.

En este tipo de codificadores puede ser aplicado como medida de calidad la relación señal-ruido *(SNR: Signal to noise ratio)* [3], el cual se explicará más adelante.

1.1.2 VOCODERS.- Esta técnica de codificación de voz surge en 1939 cuando Homer Dudley de los laboratorios *Bell* demostró en la *New York World's Fair* el primer Vocoder, en el cual la idea central era analizar la voz para extraer una serie de características que serían enviadas por el emisor y una vez que el receptor tuviera estas características reconstruirían la señal de voz original.

Este tipo de codificación es también conocida como codificadores paramétricos, donde se asume que la señal se voz puede ser generada por un modelo el cual es controlado por ciertos parámetros [1], por lo que los Vocoder aprovechan las características de la señal de voz para realizar una codificación más eficiente basándose en el modelo de producción de voz, el cual se muestra en la figura 1.1.

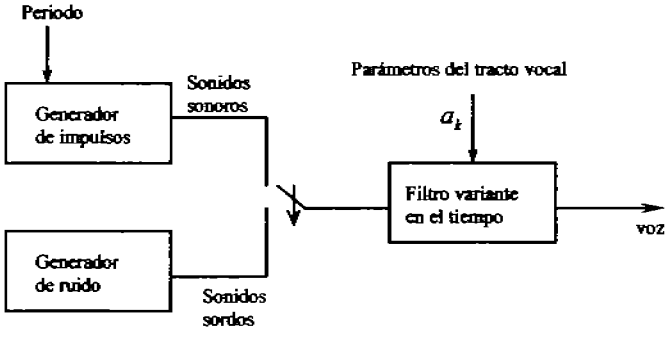


Figura 1.1. Modelo de producción de voz.

Los Vocoders independientemente de que si la forma de onda se parece a la original, intentan producir una señal que suene como la original extrayendo los parámetros del tracto vocal y la frecuencia de excitación en el transmisor. Esta información se envía al receptor donde se sintetiza la voz, produciendo una señal de voz con bajo *bit rate*, pero sonando poco natural. Por lo regular estos codificadores son utilizados para aplicaciones de seguridad (ejercito) y sus *bit rates* van de 2.4 kbps a 4.8 kbsp.

Los conceptos más importantes que obtenemos de este tipo de codificadores son además del modelo de producción de voz antes mencionado: el análisis de predicción lineal del cual se obtienen los parámetros necesarios para el filtro variante en el tiempo, los algoritmos para la obtención de la frecuencia fundamental o periodo y el tipo de sonidos en la voz, es decir sonoro o sordo todos estos mostrados en la figura 1.1. y explicando a detalle posteriormente ya que serán aplicados posteriormente.

1.1.3 Codificadores híbridos.- Los codificadores llamados híbridos combinan las técnicas de los codificadores de onda con la de los codificadores paramétricos, obteniendo la alta calidad de voz de los primeros y los bajos *bit rates* de los segundos, por lo que en su mayoría están basados en el esquema de predicción lineal.

También son conocidos como codificadores de análisis por síntesis ya que en el emisor se lleva a cabo un análisis en el cual se obtienen los parámetros de la señal para luego sintetizarla y conseguir el mayor parecido a la original.

La diferencia entre un codificador híbrido y uno paramétrico desde una perspectiva técnica esta en la forma de representar la señal de excitación para el modelo de producción de voz al eliminar la redundancia en los parámetros que son extraídos [3].

Tipo	Codificadores de onda	VOCODERS	Codificadores Híbridos
Codificadores	DM	LPC	RELP
	PCM	LPC10	MPLPC
	DPCM		CELP
	ADPCM		VSELP
			REP-LTP
Velocidad en kbps	32 a 64	2.4 a 4.8	5.6 a 16
Valor MOS	4 a 4.5	2.5 a 3.5	3.5 a 4
Aplicación	Red telefónica conmutada	Seguridad	Comunicaciones móviles

Tabla 1.1 Clasificación de los tipos de codificación

En la tabla 1.1 podemos observar un resumen de las diferentes técnicas de codificación con sus respectivas aplicaciones, velocidades de transmisión y valor MOS.

1.1.4 Modelo de producción de voz.- La voz es básicamente generada como una onda acústica que es radiada desde los orificios nasales y la boca cuando el aire es expulsado de los pulmones dando como resultado del flujo de aire perturbado por la construcción interna del cuerpo, con base en esto es muy útil interpretar a la producción de voz como un filtro acústico.

Existen tres cavidades centrales en el sistema de producción de voz: la nasal, la oral y la faringe. El tracto pulmonar o respiratorio a su vez esta formado por los pulmones y la tráquea (nasal) el cual tiene la función de generar el flujo de aire que pasará a través de la laringe donde se generarán los sonidos para finalmente llegar al tracto vocal, el cual esta formado por la faringe y la cavidad oral y modular dichos sonidos produciendo la voz que escuchamos.

Cuando llegamos a la función de la laringe, es decir a la creación de de sonidos, estos sonidos o excitación adquieren ciertas características que se clasifican como: fonación, susurreo, fricación, compresión y vibración **[4]**.

Fonación. Se refiere a la oscilación de las cuerdas vocales por los movimientos de los cartílagos. La apertura y cierre de las cuerdas secciona el pulso de aire en pulsos cuasi periódicos llamados pulsos glotales, con una frecuencia fundamental llamada tono (*pitch*).

Susurreo. Es generado en la laringe. Las cuerdas vocales están juntas, pero en lugar de sellar completamente la glotis existe una pequeña abertura triangular entre estos cartílagos. El aire que corre a través de esta apertura genera turbulencias, que ocasionan ruido de banda ancha, el cual sirve como señal excitadora.

Fricación. Es similar al susurreo en cuanto al aire turbulento que genera ruido de banda ancha, pero existe un lugar de articulación adicional en el tracto vocal. Dado que el lugar de articulación es cerca de los labios, sólo una pequeña parte del tracto vocal está entre las fuentes de excitación y el aire de salida. La fricación se da con las letras "f" y "s".

Compresión. Cuando el tracto vocal está prácticamente cerrado y una persona sigue exhalando, la presión aumenta y resulta un pequeño transitorio. La combinación de un silencio pequeño seguido por una ráfaga de ruido crea una excitación aperiódica, la onda de presión es una función escalón con un aspecto inverso a la frecuencia. La compresión se da con las letras "p" y "j".

Vibración. Es cuasi periódica y puede ocurrir en muchos lugares del tracto vocal, su efecto principal es la interrupción o una modulación semejante a la fonación, que sea rápida y repetitiva.

El modelo y forma del tracto vocal y nasal cambia continuamente en el tiempo, creando un filtro acústico variante en el tiempo **[3]**.

Con base en la figura 1.1 el filtro variable en el tiempo tiene dos posibles señales de entrada que dependerán del tipo de señal, sonora o no sonora. Para las señales sonoras, la excitación será un tren de impulsos con determinada frecuencia fundamental (*pitch*), mientras que para las señales no sonoras, la excitación será un ruido aleatorio. La combinación de estas dos señales modela el funcionamiento de la glotis.

El espectro de frecuencias para la señal vocal se puede obtener a partir del producto del espectro de la excitación con la respuesta en frecuencia del filtro. En el tracto vocal muestra muchas resonancias; sin embargo, se consideran solo las tres o cuatro primeras, mismas que toman el nombre de frecuencias formantes y cubren un rango de frecuencias que oscila entre los 100 y 3500 Hz. Esto se debe a que las resonancias de alta frecuencia son atenuadas por la característica frecuencial del tracto, que tiende a actuar como un filtro paso bajas, con una caída de aproximadamente -12 dB por octava **[5]**.

1.1.5 Procesado de señales de voz.- Para poder facilitar el proceso de extracción de características de la señal de voz y sobretodo para poder aplicar correctamente el método de predicción lineal que se basan la mayoría de las técnicas de codificación, es necesario aplicar a la señal de voz de entrada un filtro de pre-énfasis, hacer una segmentación y aplicar cierta ventana, es por eso que en esta sección se describen los fundamentos de dichos conceptos.

Filtro de pre-énfasis.- Inicialmente es necesario aplicar un filtro de pre-énfasis ya que el modelo *LPC* funcionará adecuadamente con las frecuencias bajas, pero hará un pobre trabajo con las frecuencias altas debido a que en el espectro de la voz existe una caída de - 6 [dB/octava], conforme la frecuencia aumenta. Esto se debe a la combinación de una caída de - 12 [dB/octava] ocasionada por la fuente de excitación de la voz y un incremento de + 6 [dB/octava] ocasionado por la radiación de la boca. Esto significa que, cada vez que la frecuencia aumenta al doble, la amplitud de la señal se reduce en un factor de 16. Por lo que se desea compensar esta caída de - 6 [dB/octava] con una manipulación de la señal de voz que de un incremento de + 6 [dB/octava] en el rango apropiado, de manera que la medición del espectro tenga un rango dinámico similar a lo largo de todo su ancho de banda **[6]**.

La función de transferencia de dicho filtro es $1-az^{-1}$ *(1.1)*, el cual enfatiza las altas frecuencias antes de ser procesada. El valor del coeficiente a esta alrededor de 0.9, con un valor típico de $a = {}^{15}\!/_{16} = 0.9375$.

Segmentación y ventanas.- Es importante tomar en cuenta que si se va aplicar el método de predicción lineal para la extracción de características se tome en cuenta que es necesario que la señal sea estacionaria, es decir, que su comportamiento no cambie con el tiempo y la señal de voz no es una de ellas.

Sin embargo, cuando una señal de voz se analiza en periodos de tiempo muy cortos (de 5 a 100 [ms]), la variación de sus características estadísticas es muy pequeña y se puede considerar una señal cuasi estacionaria [7]. Por lo tanto, lo que se hace es truncar la señal de voz en pequeños segmentos llamadas tramas (*frames*). Esta segmentación se realiza al multiplicar la señal de voz s(n), con una señal ventana w(n), la cual tiene la característica de que es cero fuera del intervalo que nosotros queremos extraer. Algunos ejemplos de ventanas son: rectangular, Rectangular, Bartlett (triangular), Hamming, Hanning y Blackman; se muestra cada una de sus ecuaciones y su gráfica en la figura 1.2.

- Rectangular $$w(n)=\begin{cases}1; & 0\le n\le N-1\\ 0; & c.o.v\end{cases} \qquad (1.2)$$

- Bartlett $$w(n)=\begin{cases}\dfrac{2n}{N-1} & ; \quad 0\le n\le \dfrac{N-1}{2}\\ 2-\dfrac{2n}{N-1} & ; \quad \dfrac{N-1}{2}\le n\le N-1\\ 0 & ; \quad c.o.v.\end{cases} \qquad (1.3)$$

- Hamming $$w(n)=\begin{cases}0.54-0.5\cos\left(2\pi\dfrac{n}{N-1}\right); & 0\le n\le N-1\\ 0 & ; \quad c.o.v\end{cases} \qquad (1.4)$$

- Hanning $$w(n)=\begin{cases}0.5-0.5\cos\left(2\pi\dfrac{n}{N-1}\right); & 0\le n\le N-1\\ 0 & ; \quad c.o.v\end{cases} \qquad (1.5)$$

- Blackman $$w(n)=\begin{cases}0.42-0.5\cos\left(2\pi\dfrac{n}{N-1}\right)+0.08\cos\left(2\pi\dfrac{2n}{N-1}\right); & 0\le n\le N-1\\ 0 & ; \quad c.o.v\end{cases} \qquad (1.6)$$

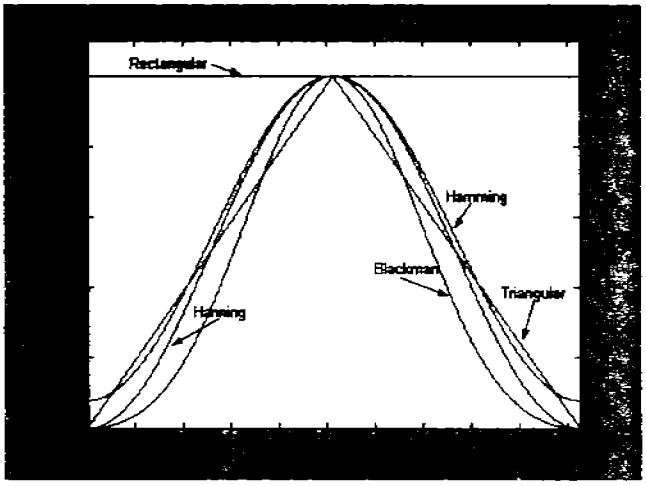


Figura 1.2. Tipos de ventanas.

En la práctica existe un compromiso para que el valor de N se encuentre entre 120-260 muestras (15 -30 ms), el cual también es determinado por razones prácticas, debido a que cuando la voz es analizada y se extraen los parámetros, estos son enviados al transmisor, el cual requeriría más alta velocidad de transmisión si las tramas fueran muy pequeñas, por lo tanto solo es necesario utilizar una longitud de muestras necesario para obtener uno o dos periodos fundamentales para cada trama [8].

En algunas aplicaciones se utiliza un traslape entre ventanas, donde la distancia entre ventanas es llamada periodo de trama (*frame period*) con un valor de entre 10-30 [ms]. El tamaño de esta dependerá de la calidad de voz, entre más pequeño sea el periodo de trama, mejor será la calidad, sin embargo esto se reflejará, como se menciono anteriormente, en la velocidad de transmisión.

1.1.6 Análisis de Predicción lineal.- Es uno de los más poderosos métodos de análisis de voz, donde el principal objetivo es modelar y remover las correlaciones entre muestras (de tiempo corto) por un muy eficiente filtro, por lo que es necesario tener un modelo teórico, que no es más que el diagrama de bloques de la figura 1.1.

En *LPC* el tracto vocal es modelado como un filtro digital todo polos, el cual se muestra en la ecuación *(1.7)*, en la cual p es el orden del modelo o del filtro.

$$H(z)=\frac{G}{1+a_1z^{-1}+\ldots+a_pz^{-p}}=\frac{S(z)}{E(z)} \tag{1.7}$$

Si definimos a s(n) como la salida del modelo y e(n) como la excitación de entrada, la ecuación *(1.7)* puede ser escrita en el dominio del tiempo como

$$s(n)=Ge(n)-a_1s(n-1)-\ldots-a_ps(n-p) \tag{1.8}$$

Así, cada muestra de voz es obtenida como una combinación lineal de las muestras previas con una contribución de la excitación.

Para completar la representación del modelo *LPC*, necesitamos obtener los coeficientes del filtro a_i que minimicen el error de predicción cuadrático medio y obtener la ganancia G que representa la energía de la señal *(1.32)*. Para lo cual, empezamos con definir el segmento de voz y el segmento de error en un tiempo *n* como

$$\begin{aligned} s_n(m)&=s(n+m)\\ e_n(m)&=e(n+m)\end{aligned} \tag{1.9}$$

Se busca minimizar la señal del error cuadrático medio en el tiempo *n*

$$E_n=\sum_m e^2(m) \tag{1.10}$$

Al usar la definición de $e_n(m)$ en términos de $s_n(m)$, se puede escribir como

$$E_n=\sum_m\left[s_n(m)-\sum_{k=1}^{p}a_ks_n(m-k)\right]^2 \tag{1.11}$$

Para resolver esta ecuación, se deriva parcialmente E_n con respecto a cada a_k y el resultado se iguala a cero.

$$\frac{\partial E_n}{\partial a_k}=0,\quad k=1,2,\ldots,p \tag{1.12}$$

Resultando:

$$\sum_{n} s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_{n} s_n(m-i)s_n(m-k) \qquad (1.13)$$

Observando que los términos de la forma $\sum_{n} s_n(m-i)s_n(m-k)$ representan las *covariancias* de los segmentos $s_n(m)$, se pueden escribir como:

$$\phi(i,k) = \sum_{n} s_n(m-i)s_n(m-k) \qquad (1.14)$$

Por lo que la ecuación *(1.14)* se puede escribir en forma más compacta como:

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \, \phi_n(i,k) \qquad (1.15)$$

Esta última ecuación describe un conjunto de p ecuaciones con p incógnitas. Y para obtener los coeficientes de predicción óptimos, se tiene que calcular $\phi_n(i,k)$ para $1 \le i \le p$ y $0 \le k \le p$ y resolver el conjunto de p ecuaciones simultáneas.

Existen varios métodos para calcular los coeficientes de predicción, algunos de estos son: *covariancia, autocorrelación, enrejado, filtro inverso, estimación espectral, máxima probabilidad* y *el de producto interno.* El más utilizado es el de autocorrelación, debido a su eficiencia computacional y a su estabilidad inherente produciendo un filtro de predicción cuyos polos se encuentran adentro del círculo unitario en el plano Z [7].

Método de autocorrelación.- Al aplicar la ventana a la señal de voz se logra determinar los límites de las sumas de las ecuaciones anteriores, por lo que se consigue expresarlas de la siguiente forma:

$$s_n(m) = s(m+n)w(m) \qquad (1.16)$$

$$E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \qquad (1.17)$$

y $\phi_n(i,k)$ es definida como:

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k), \quad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{1.18}$$

o por un cambio de variable

$$\phi_n(i,k) = r_n(i-k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{1.19}$$

Además, como la función de autocorrelación es simétrica, $r_n(-k) = r_n(k)$, las ecuaciones *LPC* pueden expresarse como:

$$\sum_{k=1}^{p} r_n(|i-k|)\hat{a}_k = r_n(i) \qquad 1 \le i \le p \tag{1.20}$$

o en forma matricial:

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \cdots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \cdots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \cdots & r_n(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \cdots & r_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix} \tag{1.21}$$

Esta matriz de orden *pxp* con los valores de autocorrelación, es una matriz *Toeplitz* (simétrica con los elementos de la diagonal principal iguales), que puede resolverse eficientemente con el uso de varios procedimientos numéricos. Uno de ellos es el algoritmo de *Levinson Durbin* [7].

Algoritmo de Levinson-Durbin.-

$$E^{(0)} = r(0) \tag{1.22}$$

$$k_i = \frac{\left[r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|) \right]}{E^{(i-1)}} \qquad para \quad 1 \le i \le p \tag{1.23}$$

$$\alpha_i^{(i)} = k_i \tag{1.24}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad para \quad 1 \le j \le i-1 \tag{1.25}$$

$$E^{(i)} = \left(1 - k_i^2\right)E^{(i-1)} \tag{1.26}$$

La solución final esta dada por:

$$a_m = coeficientes\ LPC = \alpha_m^{(p)} \tag{1.27}$$

$$k_m = coeficientes\ PARCOR \tag{1.28}$$

$$g_m = coeficientes \log area\ ratio \tag{1.29}$$

Es importante resaltar que entre los coeficientes de reflexión o *PARCOR* y los coeficientes *LPC* existe una equivalencia

$$a_i^{(i)} = k_i \tag{1.30}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \qquad \begin{matrix} i = 1,\ldots,p \\ j = 1,\ldots,i-1 \end{matrix} \tag{1.31}$$

Por último, para determinar la ganancia G que necesitamos en el modelo de síntesis o de producción de voz, la obtenemos de los parámetros *PARCOR* con la siguiente expresión:

$$G^2 = E(p) = \left(1 - k_1^2\right)\left(1 - k_2^2\right)\ldots\left(1 - k_p^2\right)r(0) \tag{1.32}$$

1.1.7 Estimación de la frecuencia fundamental.- Uno de los más importantes parámetros del análisis y codificación de voz es la obtención de la frecuencia fundamental, la cual esta relacionada directamente con un conjunto de características únicas de cada persona. Así el tiempo entre la sucesiva apertura de las cuerdas vocales se le conoce como frecuencia fundamental *(pitch)*[9].

Para hombres el rango de frecuencias fundamentales es aproximadamente entre 50 y 250 Hz (4 a 20 ms) mientras que para las mujeres es de 120 a 500 Hz (2 a 8 ms). Este periodo debe ser estimado para cada trama y el diseño de un algoritmo de estimación de periodo es demasiado compleja siendo un tema de gran interés y con mucha investigación todavía por realizar.

Algunas de las técnicas que han sido propuestas se muestran a continuación:

Método de autocorrelación.- Este método tiene su sustento en la propiedad de que si se tiene una señal con un cierto periodo y a esa señal se la aplica la función de autocorrelación, la señal resultante tendrá el mismo periodo y se podrá obtener midiendo las distancias entre máximos. Por lo que si consideramos que la trama termina en el instante m, donde la longitud de la trama esta

dada por N, el valor de la autocorrelación ec. *(1.32)* refleja la similitud entre la trama s[n] para n=m+N+1 a m, con respecto a la versión desplazada s[n-l], donde l representa un tiempo de desplazamiento. El rango del lag es seleccionado para que se cubra un rango de valores del periodo, en la práctica este rango va de l=20 a 147 (54.4 a 400 Hz o 2.5 a 18.3 ms).

$$R(l,m)=\sum_{n=m-N+1}^{m} s(n)s(n-l) \qquad (1.33)$$

Calculando los valores de la autocorrelación con la ec (*1.33*) es posible encontrar el periodo del pitch asociado con el valor más alto de la autocorrelación es decir un pico de la señal.

AMDF (Average Magnitud Diference Function).- Es un algoritmo basado en el de autocorrelación. Sin embargo el aplicar dicho método se requiere una carga computacional grande de multiplicaciones y sumas. Por esta razón *AMDF* fue desarrollado como "una función de autocorrelación de hombres pobres".

El algoritmo *AMDF* esta definido para una señal ventaneada $s(n)$, la cual no es cero entre $n=0$ y $N-1$ posteriormente es aplicada a cada trama la siguiente ecuación

$$D(k)=\sum_{n=0}^{N-1}|s(n)-s(n+k)| \qquad k=0,1,2,\ldots \qquad (1.34)$$

De esta ecuación se puede observar que el periodo del *pitch* se logra cuando $s(n)$ y $s(n+k)$ tienen la misma amplitud y por lo tanto $D(k)$ mostrará un valle en ves de un pico, por lo que existe una analogía entre estos dos algoritmos.

La clasificación de los sonidos se lleva a cabo con el análisis de un umbral de energía y la propia frecuencia fundamental con base en las propiedades de sonidos sordos o sonoros. La información de la ganancia se transmite en forma del valor cuadrático medio de cada trama.

1.1.8 Análisis por síntesis.- La estructura básica de un codificador de análisis por síntesis es mostrada en la figura 1.3, donde el filtro $\frac{1}{A(z)}$ modelo antes visto de predicción de tiempo corto con la forma $\frac{1}{A(z)}=\frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}}$, donde los coeficientes a_i son lpc y p es el orden del filtro. Por otro lado

el filtro $\frac{1}{P(z)}$ es el modelo de las correlaciones de termino largo, es decir entre tramas, con la siguiente función de transferencia

$$\frac{1}{P(z)} = \frac{1}{1 - \sum_{i=-q}^{r} b_i z^{-(D+i)}} \qquad (1.35)$$

donde D es el periodo del pitch en muestras y $\{b_i\}$ son los coeficientes de predicción de tiempo largo. El número de coeficientes varia de 1 (q=r=0) a 3 (q=r=1). El retardo D y los coeficientes pueden ser determinados de la señal de voz o de la señal residual después de remover las correlaciones de tiempo corto.

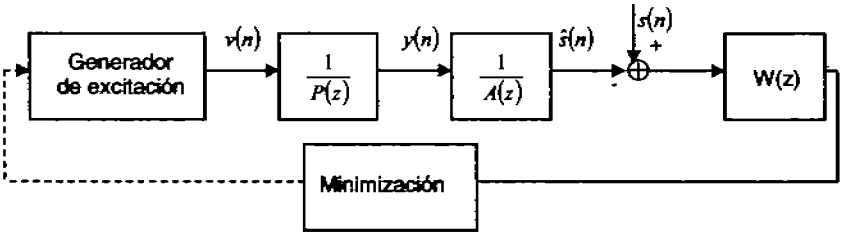


Figura 1.3. Diagrama de bloques de una estructura Análisis por Síntesis.

Una vez que los coeficientes han sido determinados, la función de excitación para el filtro es determinado de la siguiente manera: para cada L muestras de subtrama (5 a 10 ms), la excitación es determinada como el mínimo del error cuadrático medio entre la señal de voz y la señal reconstruida.

Es importante mencionar que existen muchas variaciones del esquema básico, como son el que predictor de tiempo largo puede ser omitido o se puede intercambiar el orden entre los predictores de tiempo corto y largo además de otras [10].

1.2 Cuerpos de estandarización

Para hacer eficiente la interconexión de las diferentes redes de telecomunicaciones, la estandarización de algoritmos de codificación de voz es necesaria, por lo que a continuación se muestran y describen de una manera muy general los principales cuerpos de regularización en este tema:

- *International Telecommunications Union (ITU).* Es responsable de la creación de los estándares de codificación de voz para redes de telefonía incluyendo tanto redes fijas como inalámbricas.
- *Telecommunications Industry Association (TIA).* La cual esta a cargo de la promulgación de estándares de codificación de voz para aplicaciones específicas. Es parte de la *American National Standards Institute (ANSI).*
- *European Telecomunications Standards Institute (ETSI).* Esta organización opera en Europa y centralmente se trata de una organización de manufactureros y el más importante grupo de esta organización es el Groupe Speciale Mobile (GSM), el cual tiene prominentes estándares como el RPE-LTP estándar.
- *United States Department of Defense (DoD).* Su principal función es la creación de estándares de codificación de voz, principalmente en aplicaciones militares.
- *Research and Development Center for Ratio Systems of Japan (RCR).* Es la encargada de la crear estándares para la telefonía celular japonesa.[3]

1.3 Requerimientos de diseño

El diseño y la capacidad de un algoritmo de codificación en particular esta establecido por el tipo de aplicación. En especial para la codificación de voz existen dos factores que están directamente relaciones y por decirlo de alguna manera en conflicto uno con el otro, estos factores son la calidad de voz y la velocidad de transmisión (*bit rate).* Así entre menor *bit rate* la calidad de voz se ve afectada, como ejemplo tenemos los Vocoders, sin embargo para sistemas de telefonía actuales los requisitos en la calidad de voz son estrictos y deben seguir los requerimientos de los cuerpos de estandarización.

En general para el diseño, los atributos de un codificador de voz pueden ser descritos en términos de cuatro clases: la velocidad de transmisión en bps (*bit rate*), complejidad, retardo y calidad.

1.3.1 Velocidad de transmisión en bits por segundo (bit rate).- Podemos definir al *bit rate* como el ancho de banda requerido para transmitir la voz codificada. Para las aplicaciones actuales es deseable que sea lo más baja posible para hacer más eficiente el sistema sin embargo este requerimiento como ya se había mencionado esta en conflicto con la calidad. Los sistemas de telefonía celular operan con una velocidad de transmisión de 6.7 a 13 kbps.

1.3.2 Complejidad.- La complejidad se refiere al costo computacional del algoritmo. Para la mayoría de las aplicaciones, los codificadores de voz son implementados en dispositivos de propósito especial como los DSP o de propósito general como las PC. En todos los casos para ser prácticos, el costo asociado a su implementación debe ser bajo y esto se traduce a que el número

de millones de instrucciones por segundo así como la memoria necesitada para soportar su operación debe ser lo más baja posible.

1.3.3 Retardo.- Este se refiere al retardo causado en la comunicación por el codificador. Esta definición no considera factores exteriores como el retardo por las comunicaciones a distancia o el equipo ya que dichos retardos están fuera de las manos del diseñador [10].

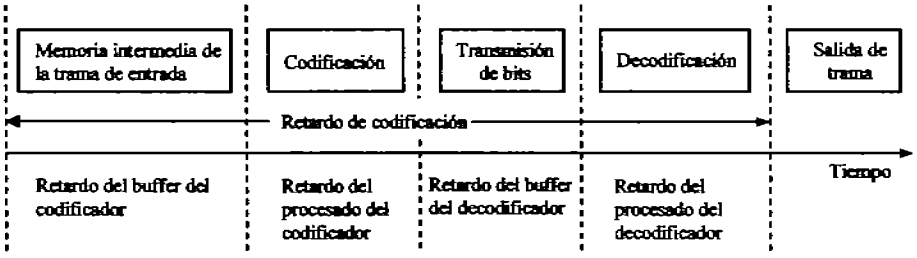


Figura 1.4. Componentes del retardo de codificación

En la figura 1.4 se muestran los componentes del retardo y donde el retardo del buffer del codificador se refiere a las muestras necesarias que se deben de almacenar por algunos codificadores antes de procesar la trama actual, esto sucede con los codificadores basados en predicción lineal.

El retardo del procesado del codificador simplemente es el tiempo que consume el codificador en procesar los datos, este tiempo puede ser reducido si se incrementa la potencia computacional y haciendo más eficiente el algoritmo.

El retardo de transmisión o retardo del buffer del decodificador es definido como el tiempo que el decodificador debe esperar para recolectar todos los bits de una trama en particular para poder empezar con el proceso de decodificación.

Por último, el retardo de procesamiento del decodificador es el tiempo requerido por el decodificador para producir una trama de la voz sintética.

1.3.4 Calidad.- La calidad es uno de los más importantes características de un codificador de voz y se refiere a un conjunto de atributos que son importantes en la percepción de la calidad, como que la voz sintética sea fácilmente entendible (inteligibilidad), que sea natural y agradable al oído, es decir que no muestre distorsión, ecos, ruido, que suene lo menos sintética y por último que la voz sea reconocible, es decir, que el escucha pueda reconocer quien es el hablante.

1.4 Métodos de evaluación de calidad

La tecnología del habla ha alcanzado un elevado grado de inteligibilidad que hacen pensar que pronto se podrá obtener niveles propios del habla natural, sin embargo no se han alcanzado dichos niveles con los *bit rates* necesarios en telefonía celular, por lo que uno de los principales puntos en los codificadores de voz es la calidad de la señal sintetizada y poder clasificar la voz recuperada en aceptable o no aceptable y al mismo tiempo poder hacer una comparación entre los diferentes tipos de codificadores con lo que respecta a calidad de la señal de voz.

Todo lo anterior nos lleva a preguntarnos por dónde se encuentra la frontera entre lo aceptable y lo no aceptable, pero sobretodo como se toma esa decisión, por lo que se ha estudiado si es posible determinar la calidad de una señal de voz en forma objetiva y repetible y si la eficiencia de mejora respecto a otra debe ser medible de forma objetiva, para poder comparar y avanzar en el desarrollo de mejores soluciones.

Por otro lado, en el diseño y evaluación de sistemas de mejora de voz, en realidad, nosotros podemos tener la señal limpia original haciendo comparaciones objetivas entre la señal procesada y la señal limpia original y será de forma irrefutable como medidor de la calidad del sistema de mejora [11].

Como hemos visto los procesos que involucran señales audiovisuales se pueden medir con dos tipos de parámetros: objetivos y subjetivos

1.4.1 Medidas Objetivas.- Este tipo de medición involucra un análisis matemático y en la práctica se utiliza la medida denominada relación señal a ruido SNR (*Signal to Noise Ratio*), la cual se define como

$$SNR = 10\log_{10}\left(\frac{\sum_{muestras} x(n)^2}{\sum_{muestras}(x(n)-y(n))^2}\right) \qquad (1.36)$$

donde la señal original se muestra como $x(n)$ y la señal sintética es $y(n)$ para todo el rango de muestras n.

Otra manera de comparar objetivamente los archivos de voz es utilizar la relación señal ruido segmental (*SSNR*: *Segmental Signal to Noise Ratio*), que promedia la relación señal ruido sobre

segmentos cortos de la forma de onda de la voz, este tipo de medida se crea para manejar la naturaleza dinámica al no ser estacionaria y matemáticamente tiene la siguiente expresión

$$SSNR = \frac{1}{N}\sum_{i=1}^{N} SNR_i \qquad (1.37)$$

donde N es el número de tramas y SNR_i es la relación señal ruido de la trama i. **[12]**

Con esta relación es posible comparar la semejanza entre el archivo original y el decodificado, de forma que si son muy parecidos, el término del denominador es muy pequeño, por lo que el valor final será muy alto. El valor de esta relación depende también de la energía del archivo de voz origen, por lo que será más alta cuanto mayor sea su energía.

La medición objetiva no siempre es relevante, inclusive esta tipo de medidas solo son significativas para los codificadores de onda, sobretodo por que estas medidas son muy sensibles a la forma de la onda y a las distorsiones de fase, las cuales no son relevantes en la percepción y sobre todo porque los Vocoders y los codificadores híbridos no preservan la forma de la señal original **[3]**.

1.4.2 Medidas subjetivas.- Con la constante evolución en el mundo de las telecomunicaciones la naturalidad y los niveles esperados por el usuario han cambiado y como consecuencia una variedad de pruebas subjetivas han surgido. La diversidad de las pruebas también ha sido modificadas por la persona que realiza la prueba y que encaje con sus propias circunstancias y objetivos, lo que por otro lado, se ha creado la necesidad de estandarizar una metodología para poder comparar los resultados de las pruebas en cualquier lugar y diferentes tiempos **[13]**, uno de los métodos de estandarización para medir la calidad en señales audiovisuales es presentados en este punto.

Este tipo de pruebas consiste en que a un grupo de personas escuchan algunos archivos de voz sintética y la comparan con la señal de voz original dando una calificación de como se escucha, posteriormente se saca el promedio de dichas calificaciones. Las pruebas son normalmente hachas bajo ciertas condiciones, algunas de las más importantes pruebas son:

- Calificación promedio de opinión (*MOS:Mean Opinion Score*)
- Tests de rimas (*RT: Rhyme Test*)

Calificación promedio de opinión (MOS: Mean Opinion Score).- Fue desarrollado por laboratorios Bell definiendo el valor como la medición de calidad estadística de la reproducción de la voz respecto a

varios tipos de codificadores. El concepto principal de esta prueba, es que la calidad de la salida de un codificador – decodificador es juzgada por un considerable número de escuchas, los cuales asignan un valor a la señal de voz después de comparar la señal original con la salida del sistema codificador – decodificado, este valor va de 1 a 5 con las siguientes características:

Valor MOS	Calidad de voz	Esfuerzo del escucha	Nivel de distorsión
1	Mala	No es entendible aplicando un esfuerzo viable	Muy ruidoso y objetable
2	Pobre	Se requiere un esfuerzo considerable	Ruidoso pero no objetable
3	Considerable	Un esfuerzo moderado es requerido	Perceptible, ligeramente ruidoso
4	Buena	Atención necesaria, no se requiere un esfuerzo apreciable	Apenas perceptible, no ruidoso
5	Excelente	No se requiere esfuerzo	Imperceptible

Tabla 1.2 Características de los valores MOS.

Por último, para obtener el valor *MOS*, se hace un promedio de los valores dados por los escuchas, con este valor se mide la aceptación. Sin embargo este tipo de pruebas tiene algunos inconvenientes, el primero de ellos, es que es un procedimiento costoso y consume tiempo.

La señal original debe tener un resultado de 5. Por lo que la mayoría de los sistemas tienen un resultado entre 3 y4. Un sistema codificador con un resultado arriba de 4 se puede catalogar como uno de muy alta calidad **[14]**.

Prueba de rimas.- La prueba fue diseñado por Fairbanks en 1958 y su versión más actual se conoce como Test de Rimas Modificado (*MRT, Modified Rhyme Test*). Se trata de una prueba formada por estímulos consistentes en palabras monosilábicas con la estructura consonante-vocal-consonante, en el que los escuchas deben elegir una palabra entre seis alternativas. Las palabras difieren en un único segmento, que se encuentra o en posición inicial o en posición final.

La adaptación al castellano ha sido llevada a cabo por Aguilar en 1991, manteniendo las características del prueba original en inglés: estímulos monosilábicos y estructura *CVC* (aceptando *CV* o *VC* en algunos casos). El requisito de la monosilabicidad plantea problemas importantes,

dado que en ciertos casos ha debido recurrirse a palabras poco familiares o reducirse el número de alternativas ante la imposibilidad de encontrar seis palabras que sólo difieran en la consonante inicial o en la final. Por este motivo, el número de alternativas en la respuesta se ha reducido a 4.

En el momento de seleccionar los monosílabos, se ha tenido en cuenta la mayor o menor frecuencia de aparición de la consonante, tendiendo a una aparición proporcional a la que se encuentra en la lengua; aún así, el equilibrio fonético no es un requisito de la Prueba de Rimas.

En la tabla 1.3 se presenta algunos ejemplos de los estímulos de la prueba, para realizarlo, el oyente debe señalar únicamente cuál es la palabra que oye entre todas las de la serie.

	A	B	C	D
1	Van	Vas	Bah	Bar
2	Ved	Ven	Ves	Ver
3	Dad	Dan	Dar	Das
4	Sol	Son	Sor	Sos

Tabla 1.3. Ejemplo de Prueba de rimas modificada.

Existen además otras herramientas, entre las que citaremos el *Diagnostic Rhyme Test* (*DRT*) de Voicers (1984), adaptado al castellano por Nadeu (1987) y actualmente en curso de revisión.

También se dispone de una versión castellana del *Fast Diagnostic Test* (*FDT*), originalmente concebido por Loman y Van Beezoijen (1988), en el que se contempla la inteligibilidad de todas las combinaciones posibles de consonante vocal en palabras de estructura *CVC* y *VCV* [15].

Los resultados de esta encuesta son reportados como el porcentaje de respuestas correctas con un ajuste de incógnitas. El rango de posibles valores es de 0 a 100 % y es calculado como:

$$DRT = \frac{correctas - incorrectas}{total} x100 \qquad (1.38)$$

CAPÍTULO 2
CODIFICACIÓN DE VOZ EN GSM

El servicio más importante que ofrece GSM al usuario es la transmisión de voz y el requerimiento técnico general es simple: transmitir señales de voz con un nivel aceptable de calidad minimizando el número de bits que se necesita transmitir, por lo que para lograr este objetivo existen tres algoritmos de codificación que son usados en los sistemas de comunicación GSM, los cuales son: Codificación VSELP (*Vector Sum Excited Linear Predictive*), Enhanced Full-Rate y RPE-LTP (*Pulse Excitation – Long Term Prediction)* cada uno con diferentes características técnicas, capacidades y desarrollo comercial que serán mostradas en este capítulo.

2.1 Codificación VSELP (Vector Sum Excited Linear Predictive)

El esquema de *VSELP* conocido como el estándar *Half-Rate GSM* **[16]** se muestra en la figura 2.1, donde existen 4 diferentes modos de operación y diferentes diagramas de bloques para dichos modos. Los coeficientes del filtro de síntesis son determinados cada 20 ms con el orden de predicción lineal de 10 (p=10), este intervalo es dividido en cuatro subtramas de 5 ms para la optimización de la excitación.

A cada uno de los cuatro modos de síntesis antes mencionados, les corresponde diferentes modos de excitación, con lo cual implica una determinación de diferentes grados de sonoridad en la señal de voz, esta determinación esta basada en la ganancia *LTP*, la cual es típicamente alta para segmentos sonoros altamente correlacionados y baja para segmentos sordos no correlacionados como el ruido.

En el modo 0 o sordo, la señal de voz es sintetizada por la superposición de las dos salidas escaladas por las ganancias G_1 y G_2 de las bibliotecas entrenadas *(trained codebooks)* que consta de 128 entradas para generar la señal de excitación, la cual es filtrada por A(z) y un post - filtro espectral.

En los modos 1, 2 y 3 donde la entrada de voz tiene algún grado de sonoridad, la excitación es ahora generada por la superposición de una biblioteca entrenada *(trained codebook)* de 512 entradas escalado por la ganancia G_3 y una biblioteca adaptiva escalada por la ganancia G_4.

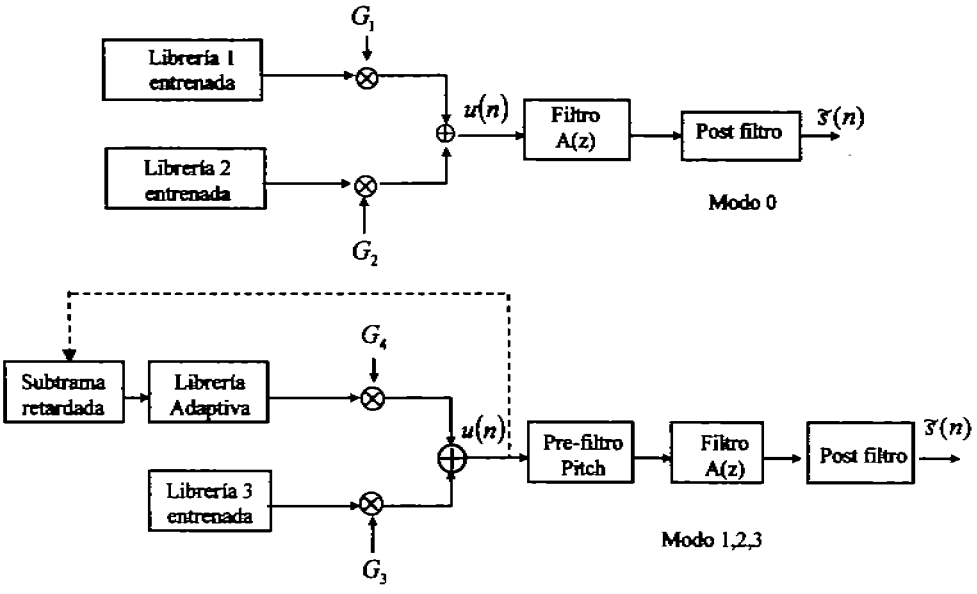


Figura 2.1. Esquema del codificador VSELP half-rate GSM, para todos sus modos

En el codificador VSELP, los 10 coeficientes de reflexión codificados para cada trama son agrupados en tres vectores $v_1 = [r_1 r_2 r_3]$, $v_2 = [r_4 r_5 r_6]$ y $v_3 = [r_7 r_8 r_9 r_{10}]$. Los vectores son cuantizados secuencialmente, de v_1 a v_3, usando un *VQ codebook* (C_i) b_i - bit para v_i, donde b_i, $i = 1,2,3$ son 11,9 y 8 b respectivamente. El vector v_i es cuantizado para minimizar la energía del error de predicción del *jth* estado del filtro lattice donde r_j es el orden más alto del coeficiente en el vector v_i. La complejidad computacional asociada con cuantizar v_i es reducida por buscar solo un pequeño subconjunto del vector código de C_i. El subconjunto es determinado primero al buscar en un *codebook* precuantizador de tamaño c_i bits, donde $i = 1,2,3$ son 6, 5 y 4 b respectivamente.

Cada vector en el *codebook* precuantizador esta asociado con $2^{b_i - c_i}$ vectores en el *target codebook*. El subconjunto es obtenido al combinar todos lo vectores código en C_i que están

asociados con la mejor concordancia de los vectores precuantizados. De esta manera, un factor de reducción en complejidad computacional cercana a $2^{b_i - c_i}$ es obtenida para la cuantifican de v_i.

Parámetros	*Bits/trama*
Coeficientes LPC	28
Bandera de interpolación LPC	1
Modo de excitación	2
Modo 0:	
Índice de Codebook1	4x7=28
Índice de Codebook2	4x7=28
Modos 1,2,3	
LTPD (subtrama 1)	8
Δ LTPD (subtrama 2,3,4)	3x4=12
Índice de Codebook 3	4x9=36
Energía de la Trama	5
Ganancia de excitación	4x5=20
Total no. De bits	112bits/20ms
Bitrate	5.6 [Kbps]

Tabla 2.1 Bitrate de VSELP Half-Rate GSM [17]

Con este tipo de codificación y con base en la tabla 2.1 obtenemos un *bitrate* de 5.6 Kbps, sin embargo la carga computacional debido a la complejidad del algoritmo es de 13.5 MIPS (millones de instrucciones por segundo). Lo cual implica un retardo considerable y por lo tanto provee un servicio regular, por lo que existen muy pocas implementaciones de este codificador.

2.2 Enhanced Full-Rate GSM.

Este codificador esta basado en el modelo de excitación ACELP y su funcionamiento empieza con la aplicación de un filtro de pre-énfasis y de nueva cuenta la cuantización espectral es realizada trama por trama mientras que la optimización de la excitación es calculada por subtramas.

Para la cuantización espectral se utiliza un análisis LPC de orden 10 el cual es realizado dos veces por trama de 20 ms, utilizando dos diferentes ventanas asimétricas de 30 ms de duración. Los coeficientes de reflexión son calculados de la señal ventaneada con el algoritmo de Levinson Durbin, posteriormente los coeficientes LPC son convertidos a LSFs y cuantizados usando SMQ (*Split Matrix Quantizer*). Este método consiste en que el término largo de LSF es quitado,

definiendo los vectores p_1^n y p_2^n para la trama n, correspondiendo a las dos ventanas antes mencionadas. Posteriormente el conjunto de n tramas LSF es predecido del conjunto LSF previamente cuantizado $\tilde{p}_2^{n-1}$, tomando en cuenta la correlación de 0.65 de sus términos largos.

Ambos LSFs vectores diferencia fungen como entrada a la SMQ. Una submatrix de *2x2* de los primeros dos LSF de ambos vectores LSF son cuantizados buscando a través de un *codebook* de 128 entradas.

Similarmente, el tercer y cuarto vector LSF de ambos LSF son cuantizados utilizando un *codebook* de 256 entradas. Finalmente, después de encontrar la mejor entrada del *codebook* para todas las submatrices de 2x2, los valores previos predecidos son sumados para producir los vectores LSF cuantizados, $\tilde{p}_1^n$ y $\tilde{p}_2^n$ respectivamente.

Con lo que respecta a la búsqueda en el *codebook* adaptivo, un análisis del pitch de lazo cerrado y abierto son usados como sigue: Con base en la voz ponderada una búsqueda del pitch en lazo abierto se lleva a cabo dos veces por cada trama de 20 ms o una vez cada dos subtramas favoreciendo valores de bajas frecuencias de muestreo con el objeto de evitar que se repita dicha frecuencia de muestreo.

Posteriormente una búsqueda de lazo cerrado para valores de frecuencia de muestreo entera es conducida sobre una base de subtramas, esto se restringe al rango [$T_0 \pm 3$] en la primera y tercer subtrama, con el objeto de mantener una baja complejidad de búsqueda. Como las subtramas segunda y cuarta el lazo cerrado de búsqueda se concentra alrededor de los valores de la frecuencia de muestreo de la subtrama previa en el rango de [-5 ... +4].

Por último los retardos del la frecuencia de muestreo fraccional se prueban también alrededor del valor de lazo cerrado en las subtramas segunda y cuarta, aun que únicamente para los retardos de la frecuencia de muestreo debajo de 95 en las subtramas primera y tercera, correspondiendo a las frecuencias de la frecuencia de muestreo en exceso de 84 HZ.

Ya determinado el lazo óptimo, la entrada del *codebook* adaptivo es únicamente identificada, mientras que su ganancia es restringida al recorrido de [0 ... 1.2] y cuantizada.

En las figuras 2.2 y 2.3 se puede observar el diagrama de bloques de los algoritmos de codificación y decodificación, así como el bitrate de 12.2 [kbps] que se logra con base en la tabla 2.2.

Parámetros	*1er. y 3er. Subtrama*	*2da y 4ta subtrama*	*No de bits*	*Total (kbps)*
Dos conjuntos de LSF			38	1.9
Ganancia de CB fijo	5	5	4x5=20	1
ACELP	35	35	4x35=20	7
Índice de CB adaptivo	9	6	2x9+2x6=30	1.5
Ganancia de CB adaptivo	4	4	16	0.8
Total			240/20ms	12.2

Tabla 2.2 Bitrate de la codificación Enhanced full – rate GSM

Por último, este tipo de codificación tiene una calidad igual e inclusive superior al codificado RPE – LTP, sin embargo esta codificación no es usada como estándar debido a la complejidad de su algoritmo que se traduce en más retrazo temporal y más utilización de recursos lo que comercialmente no es aplicado **[18]**.

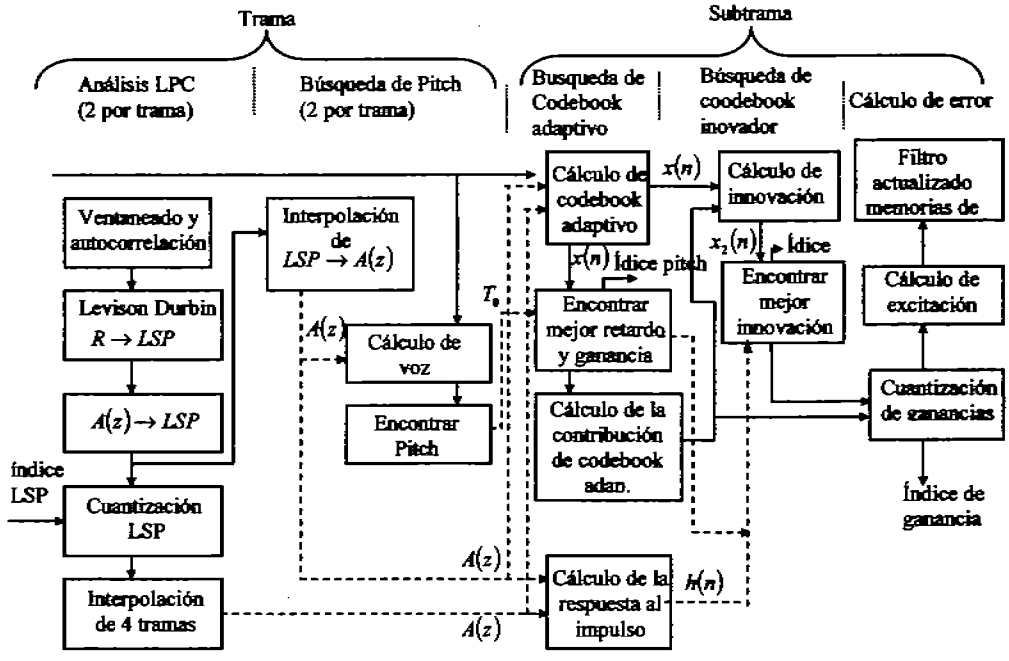


Figura 2.2. Codificador Enhanced full – rate GSM

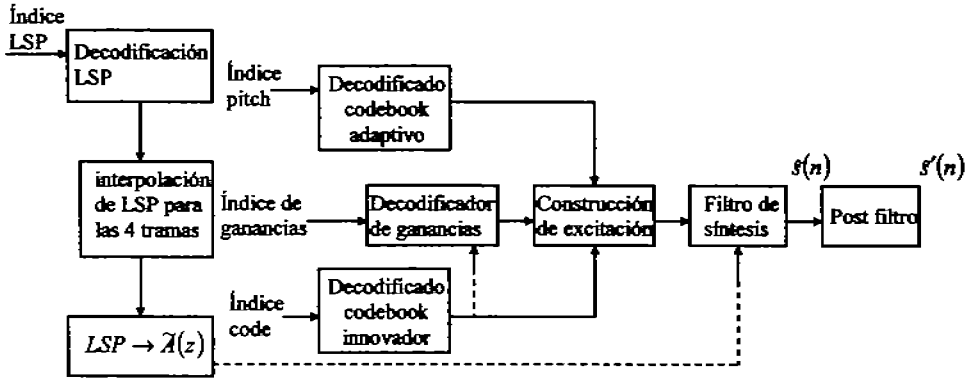


Figura 2.3 Decodificador Enhanced full – rate GSM

2.3 Codificador RPE-LTP

La codificación RPE – LTP (Regular Pulse Excitation – Long Term Prediction) es una recomendación ETSI (European Telecommunications Standards Institute) GSM 6.10 "GSM Full Rate Speech Trancoding". RPE – LTP combina las ventajas de un codificador RELP con las de un codificador MPE – LTP (Multi – Pulse Excited Long Term Prediction). Las ventajas de esta fusión, son que RELP entrega una buena calidad de voz con poca complejidad, pero hay que tomar en cuenta que la calidad de voz sigue siendo limitada debido al ruido tonal por la regeneración de altas frecuencias. Por otro lado la codificación MPE – LTP consigue una alta calidad de voz, pero tiene una complejidad alta. Así modificando RELP para incorporarle algunas características de MPE – LTP, se pudo obtener una reducción de 14.77 [Kbps] a 13 [Kbps] sin perdida en la calidad de voz. Siendo la modificación más importante la introducción del ciclo LTP.

El diagrama de bloques del codificador RPE-LTP es mostrado en la figura 2.4 donde se puede visualizar fácilmente las cuatro partes generales que lo componen: 1) Pre-procesado, 2) Análisis y filtro de retardo corto (*Short Term Prediction*), 3) Análisis y filtro de retardo largo (*Long Term Prediction*) y 4) Cálculo de RPE (*Regular Excitation Pulse*), por lo que este capítulo tiene por objetivo describir cada una de las cuatro partes antes mencionadas, definidas en el estándar actualmente utilizado, para comprender su funcionamiento.

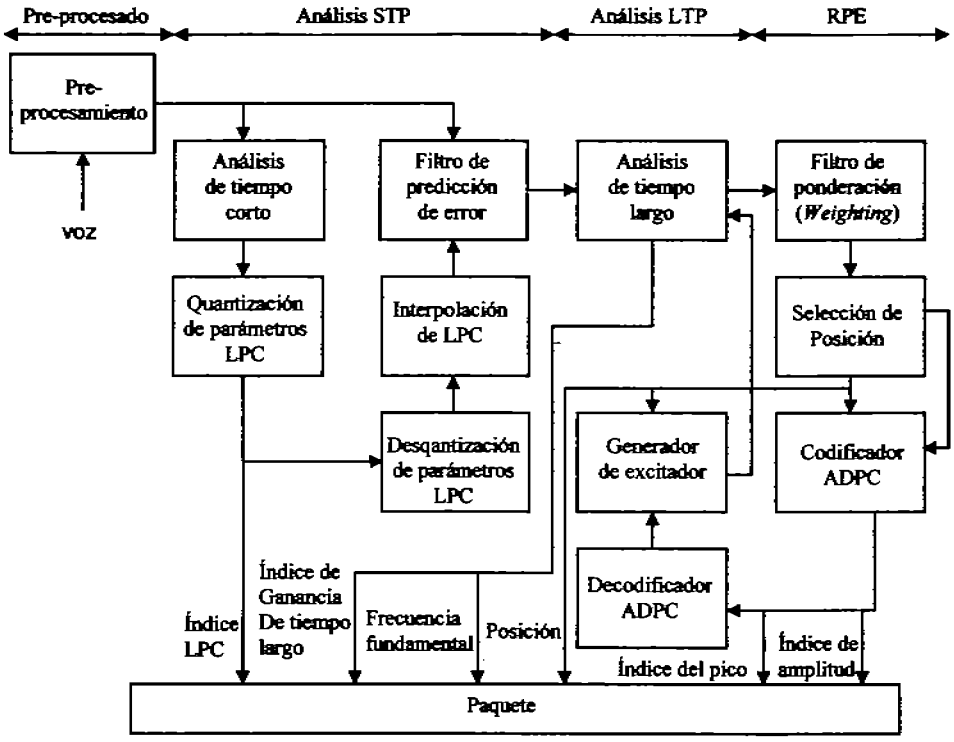


Figura 2.4. Diagrama de bloques del codificador RPE-LTP.

2.3.1 Pre – procesado.- Como principio, a la señal de voz se aplica un filtro de pre – énfasis de primer orden para incrementar la precisión numérica de los cálculos debido a la baja potencia que tienen las altas frecuencias del espectro de frecuencias de la voz. La función de transferencia que define a este filtro es la vista en el capítulo anterior ecuación (*1.1*) con la constante de c=0.9, es decir

$$H(z)=1-0.9z^{-1} \qquad (2.1)$$

De la señal pre-enfatizada se toman tramas de voz de 20 ms, que a una frecuencia de muestreo de 8000 Hz, obtenemos L=160 muestras por trama a las cuales se les aplica una ventana de Hamming, para disminuir el efecto producido en el dominio de la frecuencia por la oscilación de Gibbs, causada por el truncamiento de la señal de voz fuera de la trama analizada.

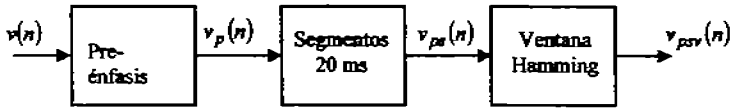


Figura 2.5. Diagrama de bloques del Pre-procesado en RPE-LTP.

La función de la ventana de Hamming es pondera por la constante c=1.5863, este último determinado por la condición de que la voz ventaneada debe tener la misma potencia que una no ventaneada **[17]**. Por lo que la ecuación resulta

$$V_{psv}(n) = v_{ps}(n) * 1.5863 * \left(0.54 - 0.46\cos 2\pi \frac{n}{L}\right) \qquad (2.2)$$

2.3.2 Predicción de termino corto (*STP: Short Term Prediction*).- A las tramas resultantes del pre – procesamiento se les aplica un análisis LPC de orden $p = 8$, por lo que como primer paso se obtienen $p+1$ coeficientes de autocorrelación para cada trama a partir de

$$R(k) = \sum_{n=0}^{L-1-k} v_{psv}(k) v_{psv}(n+k) \qquad k = 0 \ldots 8 \qquad (2.3)$$

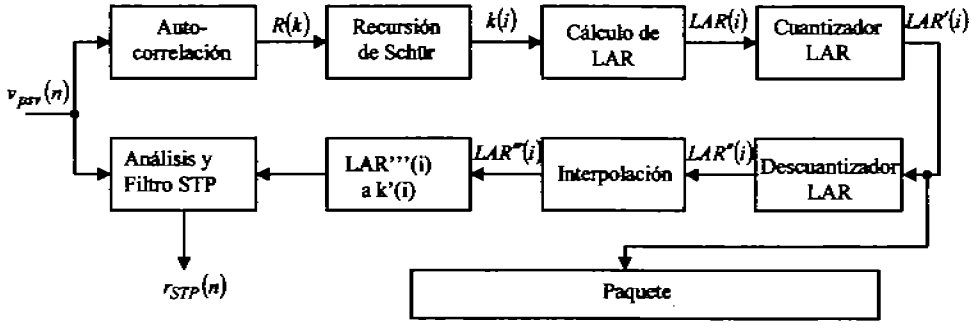


Figura 2.6. Diagrama de bloques de la Predicción de tiempo corto en RPT-LTP.

Una vez que obtenemos los nuevos coeficientes de autocorrelación podemos calcular los coeficientes de reflexión k_i utilizando la recursión de Schür, el cual es un método equivalente al

algoritmo Levinson Durbin con la diferencia de que con Schür obtenemos solamente los coeficientes de reflexión con el siguiente método de obtención [19]:

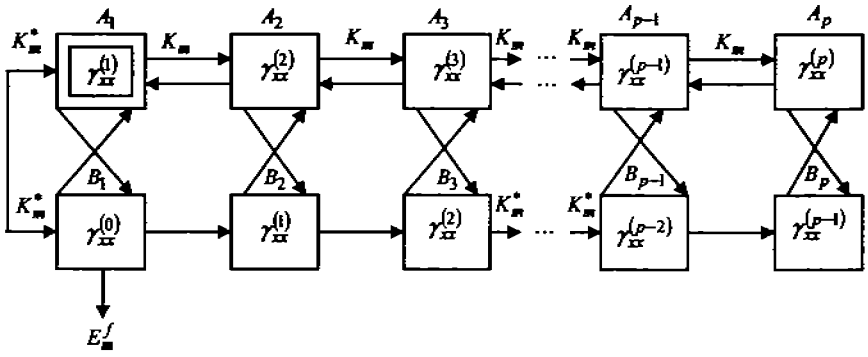


Figura 2.7. Implementación del algoritmo de Schür (Pipelined).

La implementación del algoritmo de Schür se muestra en la figura (*2.7*), el cual consiste en un proceso en cascada de p etapas de tipo lattice, donde cada etapa de dos elementos de procesado: A_1, A_2,..., A_p y B_1, B_2,..., B_p.

Inicialmente los A's son los elementos del primer renglón de la matriz generadora $G_0 = \begin{bmatrix} 0 & \gamma_{xx}(1) & \gamma_{xx}(2) & \cdots & \gamma_{xx}(p) \\ \gamma_{xx}(0) & \gamma_{xx}(1) & \gamma_{xx}(2) & \cdots & \gamma_{xx}(p-1) \end{bmatrix}$ como se ilustra en la figura (*2.7*). Los elementos B's toman el valor del segundo renglón, el proceso de cálculo de los coeficientes de reflexión empieza con la división de A_1 y B_1 con lo que se obtiene el primer coeficiente de reflexión, es decir $K_1 = -\gamma_{xx}(1)/\gamma_{xx}(0)$. El valor de K_1 es enviado al mismo tiempo a todos los A's y B's.

El segundo paso en el cálculo es actualizar el contenido de todos los elementos simultáneamente. El contenido de los elementos superiores e inferiores son actualizados como:

$$A_m: \quad A_m \leftarrow A_m + K_1 B_m, \qquad m = 2,3,\ldots,p \tag{2.4}$$

$$B_m: \quad B_m \leftarrow B_m + K_1^* A_m, \qquad m = 1,2,\ldots,p \tag{2.5}$$

El tercer paso involucra el movimiento del contenido de los A's un lugar hacia la derecha, por lo que

$$A_m: \quad A_{m-1} \leftarrow A_m, \qquad m = 2,3,\ldots,p \tag{2.6}$$

En este punto, A_1 contiene $\gamma_{xx}(2)+K_1\gamma_{xx}(1)$ mientras que B_1 contiene $\gamma_{xx}(0)+K_1^*\gamma_{xx}(1)$. Posteriormente A_1 esta listo para empezar el segundo ciclo y calcular en segundo coeficiente de reflexión $K_2 = -A_1/B_1$. El paso tres empieza con la división de A_1/B_1 y es repetida hasta que todos los p coeficientes de reflexión son calculados. Se observa que B_1 da el error medio cuadrado mínimo para cada iteración.

Una vez obtenido los ocho coeficientes de reflexión son convertidos a LAR (*Logarithmic area ratios*), esto debido a que tienen mejores cualidades para la cuantización y evitar tener errores en la transmisión de ellos utilizando la siguiente ecuación para la conversión:

$$LAR(i) = \log_{10}\left(\frac{1+k(i)}{1-k(i)}\right) \qquad (2.7)$$

donde una aproximación lineal con cinco subtramas es usado para simplificar la implementación en tiempo real:

$$LAR'(i) = \begin{cases} K(i), & if \quad |K(i)| < 0.675 \\ sign[K(i)][2|K(i)|-0.675], & if \quad 0.675 < |K(i)| < 0.95 \\ sign[K(i)][8|K(i)|-6.375], & if \quad 0.975 < |K(i)| < 1.0 \end{cases} \qquad (2.8)$$

Los coeficientes LAR(i) cuantizados LAR'(i) son localmente decodificados en un conjunto de LAR'' así como transmitidos al decodificador. Ahora los coeficientes de reflexión localmente decodificados son calculados por medio de LAR''(i) y convertidos en K'(i), los cuales son usados para calcular la señal residual $r_{STP}(n)$ en una estructura lattice como muestra la figura (*2.8*) y por medio de las siguientes ecuaciones[3]:

$$\begin{aligned} v_{p-1}(n) &= x(n)+k_p u_{p-1}(n-1), \\ v_{p-2}(n) &= v_{p-1}(n)+k_{p-1}u_{p-2} \\ &\vdots \\ v_1(n) &= v_2(n)+k_2 u_1(n-1), \\ y(n) &= v_1(n)+k_1 y(n-1), \\ u_1(n) &= -k_1 y(n)+y(n-1), \\ u_2(n) &= -k_2 v_1(n)+u_1(n-1), \\ &\vdots \\ u_{p-1}(n) &= -k_{p-1}v_{p-2}(n)+u_{p-2}(n-1). \end{aligned} \qquad (2.9)$$

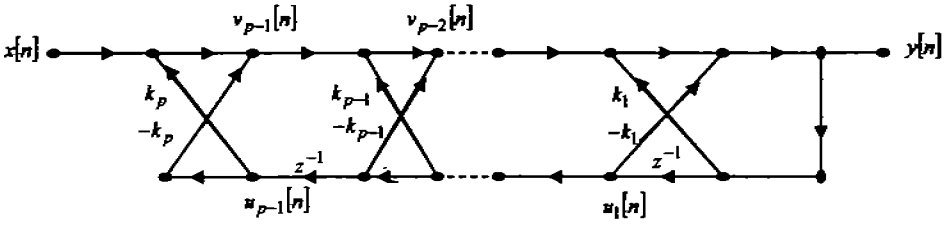


Figura 2.8.Implementación lattice del filtro lpc.

2.3.3 Predicción de termino largo (LTP: Long Term Prediction).- El objetivo principal de este modulo es remover la correlación existente entre tramas, es decir, la correlación de LTP, sin embargo, se ha demostrado que la no efectividad del predictor del LTP para remover dicha correlación, de hecho la secuencia de error de LTP es muy parecida a la del STP, conteniendo un componente fuerte de periodicidad el cual es muy cercano, si no es que, es el periodo fundamental (*pitch*) por si mismo.

Como solución a la perdida de efectividad en periodos largos, los parámetros LTP necesitan ser obtenidos más frecuentemente que los parámetros STP, por lo que experimentalmente se ha demostrado que obtener parámetros de LTP en tramas de 5 ms en vea de 20 ms incrementa la ganancia de predicción en 2.2 dB [3].

Por lo que un sistema de trama y subtrama es aplicado, el cual consiste en la simple idea de la trama de 20 ms del error de predicción del STP ($r_{STP}(n)$), es dividida en intervalos más pequeños, conocidos como subtramas, figura 2.9 y el análisis de termino largo es aplicado a cada subtrama de 5 ms separadamente.

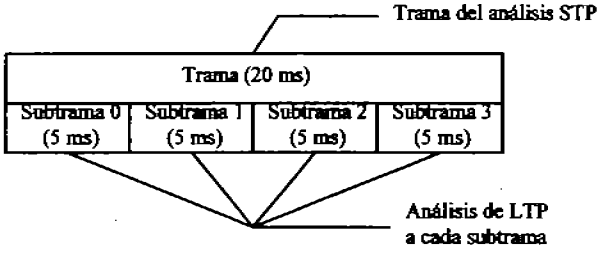


Figura 2.9. Estructura Trama – subtrama.

La minimización del error de predicción de retardo largo se consigue obteniendo un retardo d, que máximiza la correlación cruzada entre el residuo actual $r_{STP}(n)$ y los residuos previamente recibidos $r_{STP}(n-d)$. Para lograr esto las 160 muestras del residuo de retardo largo se dividen en las cuatro subtramas de 40 muestras y para cada uno de estos se calcula una correlación cruzada entre el segmento actual y las 128 muestras de residuos *STP* precedentes. El máximo de la correlación se encuentra para un retardo *d*, que con mucha probabilidad corresponde al periodo del *pitch* o a un múltiplo de éste como se menciono anteriormente. Por lo que restando el segmento altamente correlacionado multiplicado por un factor de ganancia *G*, que es la correlación cruzada normalizada para el retardo *d*, se puede eliminar buena parte de la redundancia, una vez obtenidos *G* y *d* se cuantizan con 2 y 7 bits respectivamente, llamándolos G' y d'.

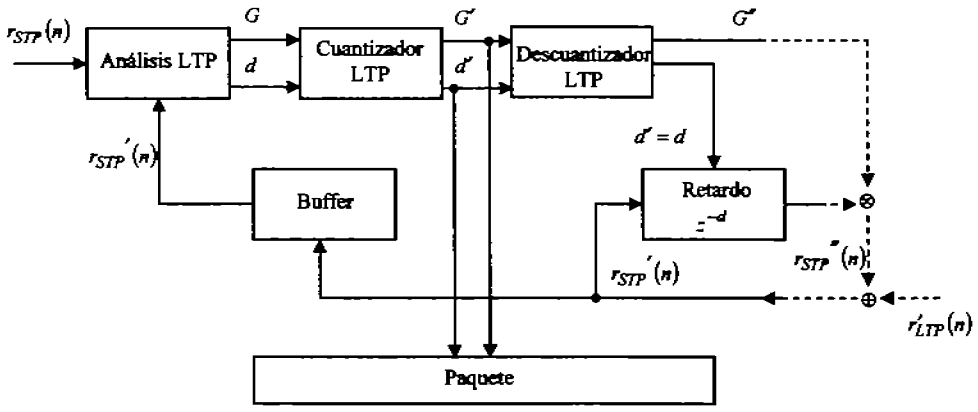


Figura 2.10. Diagrama de bloques del proceso LTP en RPE-LTP.

G' y d' son descodificados localmente para obtener G'' y d'' y poder producir el residuo localmente descodificado STP $r'_{STP}(n)$. El residuo $r_{LTP}(n)$ se calcula como la diferencia entre el residuo $r_{STP}(n)$ y la estimación $r''_{STP}(n)$, que se obtiene con los parámetros G'' y d'', siguiendo las siguientes ecuaciones:

$$r_{LTP}(n) = r_{STP}(n) - r''_{STP}(n)^{n=1...40} \qquad (2.10)$$

$$r''_{STP}(n) = G'' r'_{STP}(n-d)^{n=1...40} \qquad (2.11)$$

donde $r'_{STP}(n-d)$ representa un segmento pasado. Finalmente esa representación se actualiza usando el residuo $r'_{LTP}(n)$ y la estimación del residuo $r''_{STP}(n)$ para formar $r'_{STP}(n)$

$$r'_{STP}(n) = r'_{LTP}(n) + r''_{STP}(n)^{n=1..40} \qquad (2.12)$$

2.3.4 Bloque RPE (*Regular Pulse Excitation*).- El residuo $r_{LTP}(n)$ se pesa con un filtro paso bajas FIR con frecuencia de corte de 4 [KHz]/3=1.33 [KHz], que es esencialmente un filtro paso bajas que suaviza las variaciones entre muestras, eliminando el ruido de las altas frecuencias y haciendo la transición entre muestras más suave, con lo que obtiene una mejora en la calidad subjetiva de la señal de voz sintética [20].

Posteriormente se hace una decimación por un factor de 3. La señal residual filtrada $r_{FLTP}(n)$ se descompone en cuatro excitaciones candidatas con pulsos regularmente espaciados definidos por la ecuación (*2.13*) e ilustradas en la figura 2.11.

$$x_m(n) = x(m+3n), \qquad m = 0,1,2,3; \qquad n = 0,1,\ldots,12. \qquad (2.13)$$

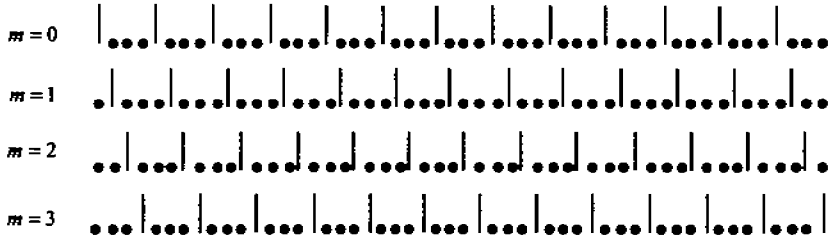


Figura 2.11. Excitaciones candidatas para L=40 y un factor de 3.

Se calcula las energías de cada secuencia decimada y la que tiene mayor energía se elige como representante del residuo LTP. Los pulsos de excitación son posteriormente normalizados a la máxima amplitud $v_{max}(k)$ de las trece muestras y son cuantizadas uniformemente con tres bits, mientras que el máximo de la amplitud se cuantiza logaritmicamente con seis bits. Como se tienen tres posibles posiciones iniciales de la secuencia decimada elegida, se necesitan dos bits para codificar la posición inicial de cada subsegmento. Esto se conoce como posición de *grid*. Las

amplitudes de los pulsos $\beta(k,i)$, la posición de *grid* y los máximos de la amplitud son localmente decodificados para obtener el residuo $r'_{LTP}(n)$, donde la secuencia decimada se rellena con ceros. El diagrama de bloques de esta sección es mostrado en la figura 2.12.

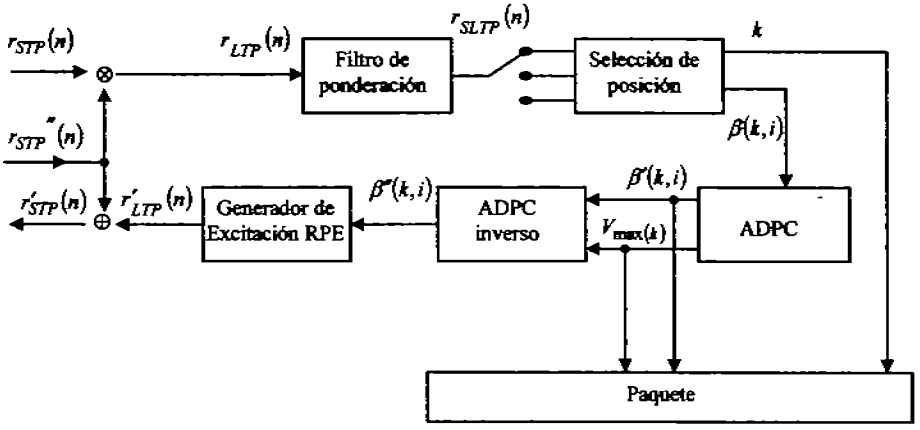


Figure 2.12. Diagrama de bloques del proceso RPE en RPE-LTP

2.4 Decodificador RPE-LTP

Con lo que respecta al decodificador RPE-LTP, en la figura 2.13 se muestra el diagrama de bloques el cual muestra el proceso inverso de los cuatro bloques descritos anteriormente: una decodificación *RPE (Regular Pulse Excitation)*, síntesis del filtro *LTP (Long Term Prediction)*, síntesis del filtro *STP (Short Term Prediction)* y por último el Post – procesado.

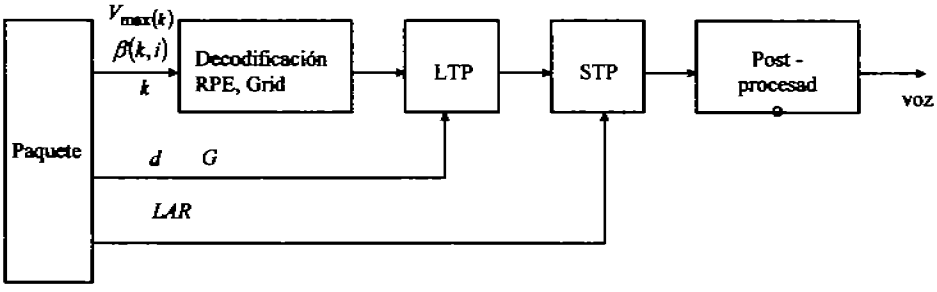


Figura 2.13. Decodificador RPE – LTP

2.4.1 Decodificación RPE.- En el decodificador, las amplitudes de los pulsos $\beta(k,i)$, la posición del grid k y los máximos de la amplitud $V_{max}(k)$ son localmente decodificados y se calculan las amplitudes de los pulsos actuales multiplicando su amplitud actual por la amplitud máxima del bloque definida por la ecuación (2.14). El modelo del residuo $r'_{LTP}(n)$ se recupera colocando apropiadamente las amplitudes de los pulsos de acuerdo con el grid inicial.

$$Amplitud(k) = V_{max}(k) * \beta(k,i) \qquad (2.14)$$

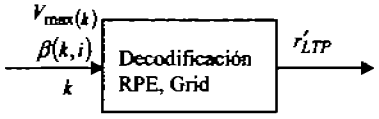


Figura 2.14. Diagrama de bloques del decodificador RPE en RPE-LTP.

2.4.2 Síntesis del filtro LTP.- Inicialmente, los parámetros del filtro LTP (G', d') son cuantizados inversamente para derivar los parámetros del filtro LTP. A continuación, el modelo de excitación recuperado $r'_{LTP}(n)$ se utiliza para excitar el filtro y recuperar un nuevo subsegmento o subtrama de 40 muestras de longitud del residuo estimado $r'_{STP}(n)$. Para realizar esto se usa la historia del residuo recuperado $r'_{STP}(n)$, retardado por d' muestras y multiplicado por G', y así se obtiene la estimación del residuo $r''_{STP}(n)$ de acuerdo con:

$$r''_{STP}(n) = G' r'_{STP}(n - d') \qquad (2.15)$$

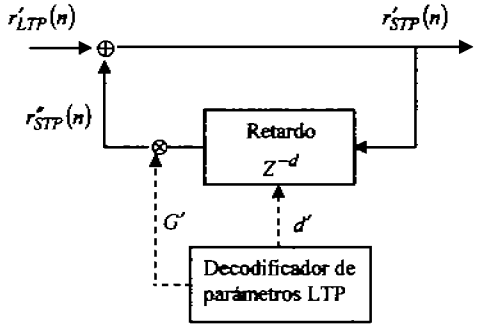


Figura 2.15. Diagrama de bloques de la síntesis de LTP en RPE-LTP.

y entonces $r''_{STP}(n)$ se usa para calcular el residuo del subsegmento recuperado más reciente con la ecuación:

$$r'_{STP}(n) = r''_{STP}(n) + r'_{LTP}(n) \tag{2.16}$$

2.4.3 Síntesis del filtro STP.- Para la síntesis del filtro STP (figura 2.17) los parámetros $LAR'(i)$ son decodificados utilizando el cuantizador inverso y convertidos a los coeficientes de reflexión por medio de la siguiente expresión

$$k(i) = \frac{10^{LAR(i)} - 1}{10^{LAR(i)} + 1} \tag{2.17}$$

Con los coeficientes de reflexión y aplicando el filtro lattice inverso de la figura 2.8 y ecuaciones *(2.9)*, ahora:

$$\begin{aligned} v_1(n) &= x(n) - k_1 x(n-1), \\ u_1(n) &= -k_1 x(n) + x(n-1), \\ v_2(n) &= v_1(n) - k_2 u_1(n-1), \\ u_2(n) &= -k_2 v_1(n) + k_2(n-1), \\ &\vdots \\ y(n) &= v_{p-1}(n) - k_p u_p(n-1). \end{aligned} \tag{2.18}$$

se obtiene la señal de voz sintética pre-énfatizada.

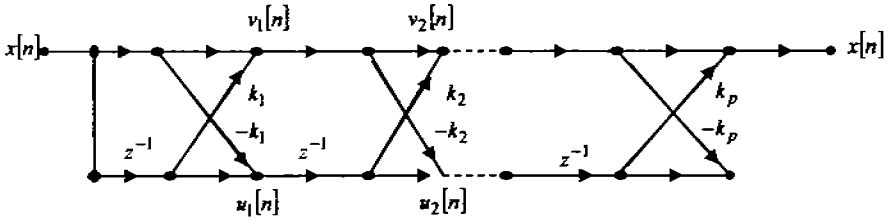


Figura 2.16. Implementación lattice del filtro inverso lpc.

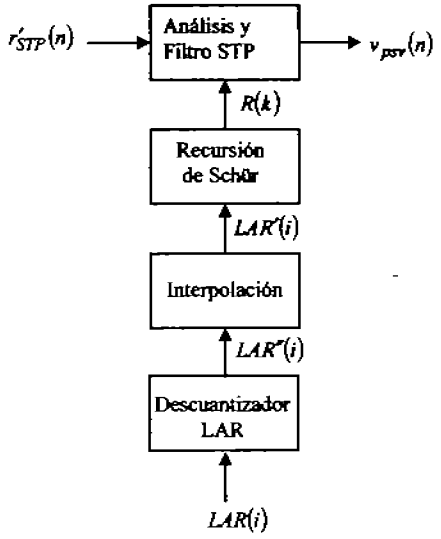


Figura 2.17. Diagrama de bloques de la síntesis de STP de RPT-LTP

2.4.4 Post-procesado.- El último bloque es el post-procesamiento y como muestra la figura 2.18 esta desarrollado por el filtro de de-énfasis, ecuación (*2.19*), que no es más que el filtro inverso de la ecuación (*2.1*).

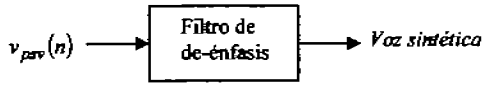


Figura 2.18. Bloque de Post-procesado en RPE-LTP

$$A(z) = \frac{1}{1 - \alpha z^{-1}} \qquad (2.19)$$

Por último la tabla 2.3 tiene el resumen de la asignación de bits para el estándar GSM 6.10 RPE-LTP. El esquema utiliza un total de 260 bits/trama, resultando una velocidad de transmisión (*bit-rate*) 13,000 bits por segundo para tramas de 20 ms y una frecuencia de muestreo de 8000 Hz, considerado robusto bajo condiciones de ruido y varios errores de transmisión por canales.

Parámetro	Número por trama	Resolución	Total de bits por trama
LPC	8	6,6,5,5,4,4,3,3	36
Periodo (Pitch) (d)	4	7	28
Ganancia LTP (G)	4	2	8
Posición (k)	4	2	8
Máximo $V_{max}(k)$	4	6	24
Amplitud $\beta(k,i)$	4*13	3	156
TOTAL			**260**

Tabla 2.3. Resumen de las asignación de bits para la codificación RPE-LTP.

También el la tabla 2.4 se presenta un resumen de las tres codificaciones presentadas en este capítulo.

Codificación de voz	**Velocidad de transmisión**	**MOS**	**MIPS [22]**	**Disponible**	**Comentarios**
Full rate	13 kbit/s	3.7	5-6	Desde los inicios de GSM	Protocolo en uso para todas las terminales producidas para datos
Half rate	6.5 kbit/s	3.8	13.5	Estándar desde la mitad de los 90's, implementación poco clara	Permite mayor capacidad de la red, calidad similar a la de full-rate, pero mas sensible con conexiones de baja calidad
Enhanced Full Rate (EFR)	13 kbit/s	4.4	18	Desde finales de 1997	Igualación en calidad de voz y menos sensibilidad en la transmisión de errores, desarrollado para PCS-1900 y muy poca implementación

Tabla 2.4. Protocolos de la codificación de voz para la familia GSM [21].

Una vez que se estableció la metodología y se definió los parámetros necesarios para el estándar actual "real" [23] del codificador de voz en telefonía celular GSM RPE-LTP, podemos establecer los posibles cambios para la mejora en la calidad de voz sin alterar la estructura y por lo tanto la velocidad de transmisión en bits/s cumpliendo con el objetivo principal de este trabajo, para lo que se tomarán las ventajas de tener diferentes métodos para la obtención de los parámetros necesarios en cada uno de los bloques descritos, una vez presentados y analizados en el siguiente capítulo se hará los cambios y las combinaciones de dichos métodos para lograr un nuevo diseño basado en el RPE-LTP sin modificar su estructura y por lo tanto sin modificar su velocidad de transmisión pero obteniendo una mejora en la calidad de voz.

CAPÍTULO 3
DISEÑOS E IMPLEMENTACIONES

3.1. Mejoras en el bloque de pre-procesado

En el capítulo anterior fueron mencionaron las características del estándar GSM, en cuanto al bloque de preprocesado, solo se establece el uso del filtro de pre-énfasis descrito en la ecuación *(2.1)* y el uso de una ventana de tipo Hamming ponderada como la describe la ecuación *(2.2)*, sin embargo es común en los codificadores de voz que como primer paso del preprocesado sea usado un filtro paso altas con función de transferencia $\left(1-z^{-1}\right)\left(1-0.999z^{-1}\right)$ para eliminar cualquier componente de DC.

Por otra parte cuando un paso en los algoritmos es la obtención de la frecuencia fundamental, a la señal de voz, comúnmente es aplicada un filtro paso bajas antes de obtener dicha frecuencia, esto es conveniente ya que la frecuencia fundamental esta relacionada con la región de las bajas frecuencias dado que su valor es menor a 500 Hz, como se había mencionado en el capítulo 1, por lo que el uso de este tipo de filtro eliminará la interferencia de los componentes de alta frecuencia así como el ruido de la banda externa.

El resultado posteriormente se pasa por el filtro de pre-énfasis antes visto con la libertad en el valor del coeficiente, sin embargo es muy recomendable para que le filtro se mantenga estable que el coeficiente tenga un valor entre $0.8 < \alpha < 0.9$, con un valor característico de $\alpha = 0.86$, además de poder utilizar una gama de ventanas para la obtención de tramas presentadas en el capítulo 1, siendo las más utilizadas la rectangular, Hanning y la propia Hamming.

3.2. Mejoras en el análisis de tiempo corto

Los filtros con función de transferencia

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}} \qquad (3.1)$$

o

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i} \qquad (3.2)$$

Son de particular importancia para la codificación de voz. En la ecuación *(3.1)* esta representado un filtro todo polos, mientas que el filtro todo ceros tiene una función de transferencia dada por *(3.2)*, estos filtros son inversos uno del otro. La constante p es el orden de el filtro y los a_i son los coeficientes del filtro. Estos filtros aparecen en todos los codificadores de voz basados en predicción lineal, por lo que a p también se le conoce como el orden del predictor y a_i's son los coeficientes de predicción lineal.

Como se vio en el capítulo anterior el estándar *GSM* utiliza una realización de dichos filtros de tipo lattice (figura 2.8 y ecuaciones *(2.18)*.). Sin embargo existe otro tipo de realización que inclusive es mucho más común en los codificadores de voz, esta es la forma directa.

3.2.1 Realización del filtro de predicción lineal en forma directa.- Con $x(n)$ siendo la entrada al filtro y $y(n)$ la salida, la ecuación en diferencias en el dominio del tiempo que corresponde a *(3.1)* es

$$y(n) = x(n) - \sum_{i=1}^{p} a_i y(n-i) \qquad (3.3)$$

Y para *(3.2)* es

$$y(n) = x(n) + \sum_{i=1}^{p} a_i x(n-i) \qquad (3.4)$$

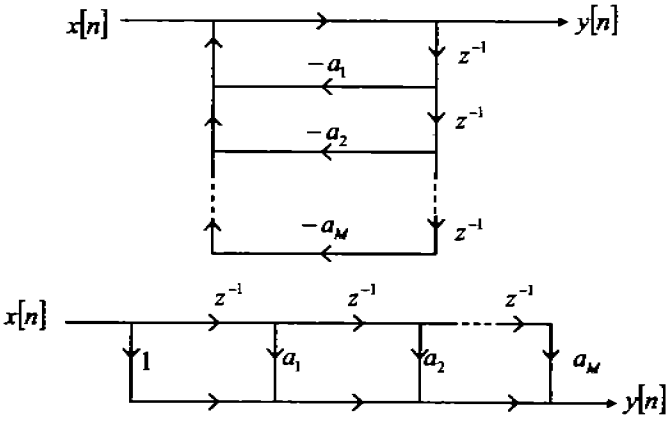


Figura 3.1 Implementación para la forma directa de un filtro todo polos (superior) y un filtro todo ceros (inferior).

En la figura 3.1 podemos ver el flujo de las señales de la ecuación *(3.1)* y *(3.2)*.

Se hace notar que la respuesta al impulso de un filtro todo polos tiene un número infinito de muestras con valores no triviales por el hecho que la versión escalada y retrazada de las muestras de salida son sumadas atrás por las muestras de entrada. Por lo que estamos hablando de un filtro con respuesta al impulso infinito (*IIR*). Para el filtro todos ceros, sin embargo, la respuesta al impulso solo tiene p+1 muestras no triviales, es decir el resto son cero, por lo que nos estaríamos refiriendo a un filtro con respuesta al impulso finito (*FIR*).

La realización de la forma directa es seguido preferida en la práctica por su simplicidad y bajo costo computacional.

Durante el análisis de predicción lineal, el método usado para resolver la matriz *Toeplitz*, ecuación *(1.21)* es normalmente el algoritmo Levinson – Durbin, donde son obtenidos tanto los coeficientes de predicción lineal (LPC o coeficientes de forma directa) como los coeficientes de reflexión. Por otro lado, también puede ser aplicado el algoritmo Leroux-Gueguen o la recursión de Schür (descrito en el capítulo anterior) donde los algoritmos solo calculan los coeficientes de reflexión. La estructura lattice permite hacer el procesamiento directamente con los coeficientes de reflexión, es decir sin convertir los coeficientes de reflexión en LPCs. Esto puede convertirse en una ventaja para los sistemas con una precisión numérica limitada debido a la perdida de precisión durante la conversión ya que esta puede generar una inestabilidad en el filtro. También, el uso de los coeficientes de reflexión nos permite tener una supervisión de estabilidad vigilando de una manera muy sencilla la condición de $|k_i| \leq 1$.

3.2.2 Algoritmo Leroux-Gueguen.- Como se menciono anteriormente los coeficientes PARCOR son obtenidos directamente a partir de los coeficientes de correlación con base en las siguientes ecuaciones **[23]**

$$\begin{aligned} e^{h+i} &= e^{h_i} + k_h + e_h^h + 1 - i \\ k_{h+1} &= -\frac{1}{e^{h_{h+1}}} e^{h_0} \\ e_0^{h+1} &= e^{h_0}\left(1 - k_{h+1}^2\right) \end{aligned} \qquad (3.5)$$

con lo cual tenemos una solución recursiva para los coeficientes k_m a partir de los valores $e^{0_i} = r(i)$.

3.3. Mejoras en el análisis de tiempo largo

Uno de los parámetros más importantes en aplicaciones de análisis, síntesis y codificación de voz es la frecuencia fundamental. Este parámetro esta directamente relacionado con el hablante ya que este es una característica única para cada persona y es estimado para cada trama, sin embargo el diseño de la estimación de la frecuencia fundamental no es trivial e inclusive es el problema más complejo a solucionar [24] y uno de los temas con más investigación relacionados con el procesamiento de voz.

La frecuencia fundamental o mejor conocida como el *pitch* por su nombre en ingles puede ser

En el capitulo 1 conocimos dos métodos para la obtención de la frecuencia fundamental, el método de autocorrelación y *AMDF*, en la cual ambos utilizan una estrategia de encontrar el pico que corresponda con la frecuencia fundamental, sin embargo estas aproximaciones pueden conducirnos a obtener valores erróneos los cuales disminuirían la calidad de voz considerablemente obteniendo un sonido demasiado sintético (ya que la calidad de la voz sintética esta altamente relacionada con una estimación precisa), estos valores erróneos probablemente corresponden a múltiplos de la frecuencia fundamental, dado que en el caso ideal, la gráfica de la autocorrelación de una señal mostrará picos en intervalos regulares separados por un periodo T, como muestra la figura 3.2, sin embargo en el mundo real, donde las señales son quasiperiodicas, como la voz, el periodo de la señal es no constante, además de tener una resolución limitada por tratarse de un sistema discreto más el ruido y la distorsión de la señal figura 3.3.

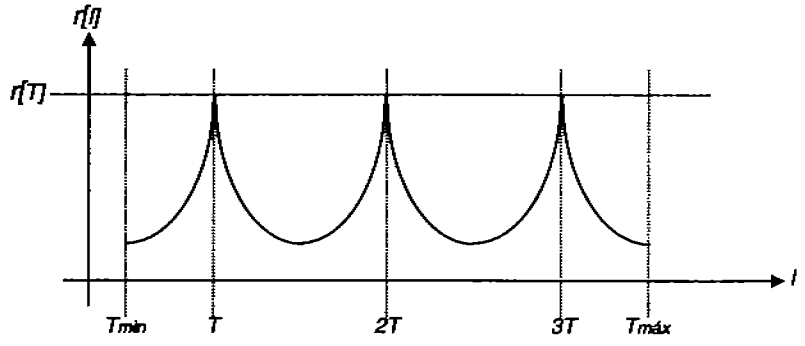


Figura 3.2 Gráfica de la autocorrelación de una señal con periodo T en un mundo ideal.

Por lo que sería de gran ayuda y mejoraría en una manera considerable si se puede analizar si la frecuencia fundamental estimada de una cierta trama es múltiplo de alguna frecuencia

fundamental, para realizar esto a cabo, un simple procedimiento es presentado para verificar dicha multiplicidad.

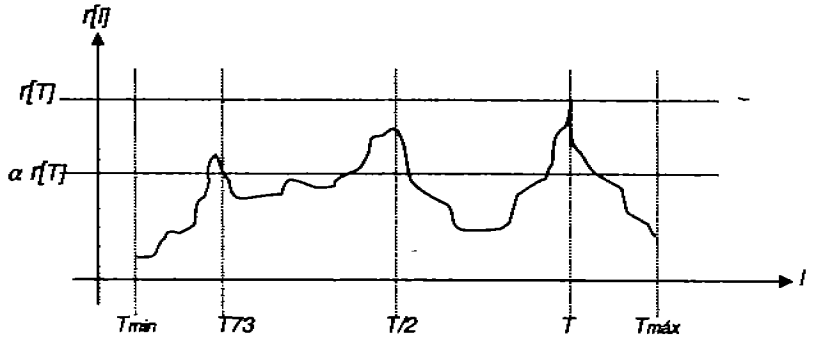


Figura 3.3 Gráfica de la autocorrelación de una señal en un mundo real.

La idea principal es verificar los valores de autocorrelación en los *lags* de $T/i \quad i=2,3,4,\cdots$, donde T es el periodo fundamental estimado.

Si se cumple que $r[T/i] > \alpha r[T]$ donde $r[\bullet]$ es la función de autocorrelación y $\alpha < 1$ es una constante positiva de escalamiento, el valor del periodo estimado se convertirá en T/i. Donde el propósito de α es hacer menor el valor del umbral de decisión, siendo necesario puesto que $r[T]$ es el pico dentro del rango buscado.

Un valor de α en el intervalo $[0.5,1]$ es una elección razonable en la práctica. La figura 3.4 muestra el diagrama de flujo del método, donde la entrada es el periodo fundamental candidato T y $r[l]$ es la función de autocorrelación. El algoritmo empieza por dividir el periodo de entrada por un rango de denominadores, controlado por i con un valor inicial de D_{min}, el cual es una constante entera determinada por el mínimo valor posible del periodo fundamental estimado. Un valor de D_{min} entre 5 y 10 es apropiado para propósitos prácticos. Puntos de prueba intermedios son encontrados al dividir T por i y redondear los resultados. Si el valor de la autocorrelación en el punto prueba es mayor que el T multiplicado por el factor de escalamiento α, entonces $T_i = redondear(T/i)$ es regresado como el periodo fundamental, donde el operador $redondear(\bullet)$ no es más que el redondeo un número al entero más cercano. Por otro lado, el

denominador i es reducido por uno y la operación es repetida hasta que se cumpla que $i < 2$. Es importante mencionar que este método puede ser usado con otras aproximaciones como el AMDF y variantes del método de autocorrelación con mínimas modificaciones.

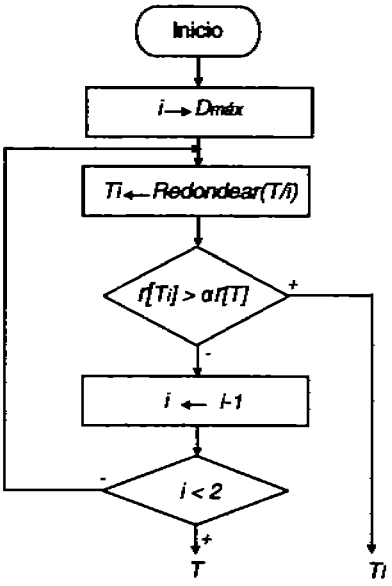


Figura 3.4 Diagrama de flujo para determinar multiplicidad de la frecuencia fundamental.

3.3.1 Método de la auto-correlación normalizada.- La señal de voz en el término largo es una señal no estacionaria por lo que los criterios antes vistos en algunos casos pueden producir errores. El criterio de normalización es similar al antes visto solo que en este método es derivado tomando en cuenta el proceso no estacionario, si tomamos en cuenta la ecuación utilizada en el método de autocorrelación $E(\tau)=\frac{1}{N}\sum_{n=0}^{N-1}[s(n)-\beta s(n+\tau)]^2$, donde β es un factor de escalamiento, o la ganancia el *pitch* y tomando la $\frac{\partial E(\tau,\beta)}{\partial \beta}=0$, dicha ganancia puede ser calculada como $\beta=\frac{\sum_{n=0}^{N-1}s(n)s(n+\tau)}{\sum_{n=0}^{N-1}s^2(n+\tau)}$ y de sustituir la ganancia óptima en la ecuación de la función error, la frecuencia

fundamental puede ser calculada de minimizar $E(\tau,\beta)=\sum_{n=0}^{N-1}s^2(n)-\frac{\left[\sum_{n=0}^{N-1}s(n)s(n+\tau)\right]^2}{\sum_{n=0}^{N-1}s^2(n+\tau)}$, lo cual es equivalente a maximizar el cuadrado de la función de autocorrelación normalizada dada por $R_n^2(\tau)=\frac{\left[\sum_{n=0}^{N-1}s(n)s(n+\tau)\right]^2}{\sum_{n=0}^{N-1}s^2(n+\tau)}$, sin embargo el uso directo de esta ecuación también puede resultar en algunos errores. Esto es porque el cuadrado de la autocorrelación puede resultar en un máximo aun si la correlación es negativa. Para eliminar este problema, se toma la raíz cuadrada del la última ecuación, la cual remueve el cuadrado de la correlación y por lo tanto elimina la posibilidad de seleccionar un lags con correlación negativa como la frecuencia fundamental. Finalmente la autocorrelación normalizada esta dada por

$$R_n(\tau)=\frac{\sum_{n=0}^{N-1}s(n)s(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}s^2(n+\tau)}} \qquad (3.6)$$

3.3.2 Función de recorte central (Centre Clipping Function).- Aunque la frecuencia fundamental de un segmento de voz sonoro puede ser estimada directamente de la señal original, el primer formante de la frecuencia puede afectar la precisión de la estimación, por lo que el espectro es aplanado para la eliminación de formantes antes de que el proceso de estimación pueda inicial. Existen dos métodos para llevar a cabo esta función, el lineal y el no lineal.

Método lineal.- Este método utiliza un filtro LPC inverso para remover los formante de la señal de voz. La principal desventaja de este método es que para una frecuencia fundamental alta, como la de un niño, el primer cero del filtro inverso puede llegar a ser la primera harmónica así como el segundo cero puede llegar a ser la segunda harmónica. Esto puede traer como consecuencia la destrucción de la periodicidad en el muestreo de señales **[25,26]**.

Método No lineal.- El especto plano en este método es conseguido por la función de recorte central, aquí se presentan tres tipos de funciones.

$$y=clip(x)=\begin{cases}x; & -CL\geq x\geq CL\\ 0; & -CL>x<CL\end{cases} \qquad (3.7)$$

$$y = clc(x) = \begin{cases} x + CL; & x \le -CL \\ x - CL; & x \ge CL \\ 0 & -CL < x < CL \end{cases} \quad (3.8)$$

$$y = \mathrm{sgn}(x) = \begin{cases} 1; & x \ge CL \\ -1; & x \le -CL \\ 0; & -CL > x < CL \end{cases} \quad (3.9)$$

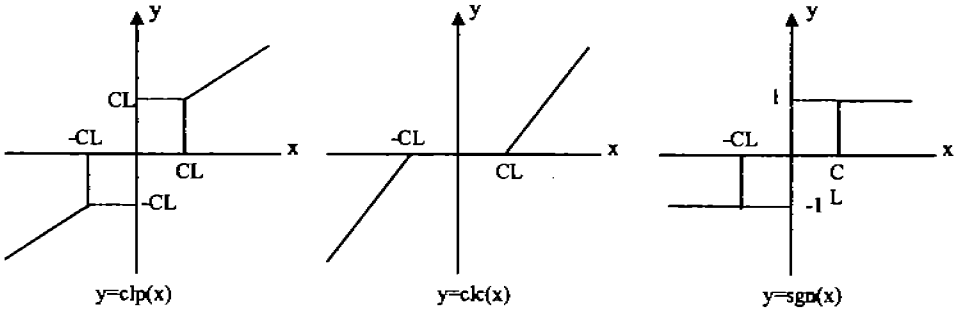


Figura 3.5 Funciones Clipper.

La señal $y(n)$ es generalmente definida como:

$$y(n) = f[s(n)].$$

La función $f[\bullet]$ puede ser cualquiera de las funciones superiores. Para el método de autocorrelación, la función "recortada" estaría definida como:

$$R_c(\tau) = \sum_{n=0}^{N-1} y(n)y(n+\tau) = \sum_{n=0}^{N-1} f[s(n)]f[s(n+\tau)] \quad (3.10)$$

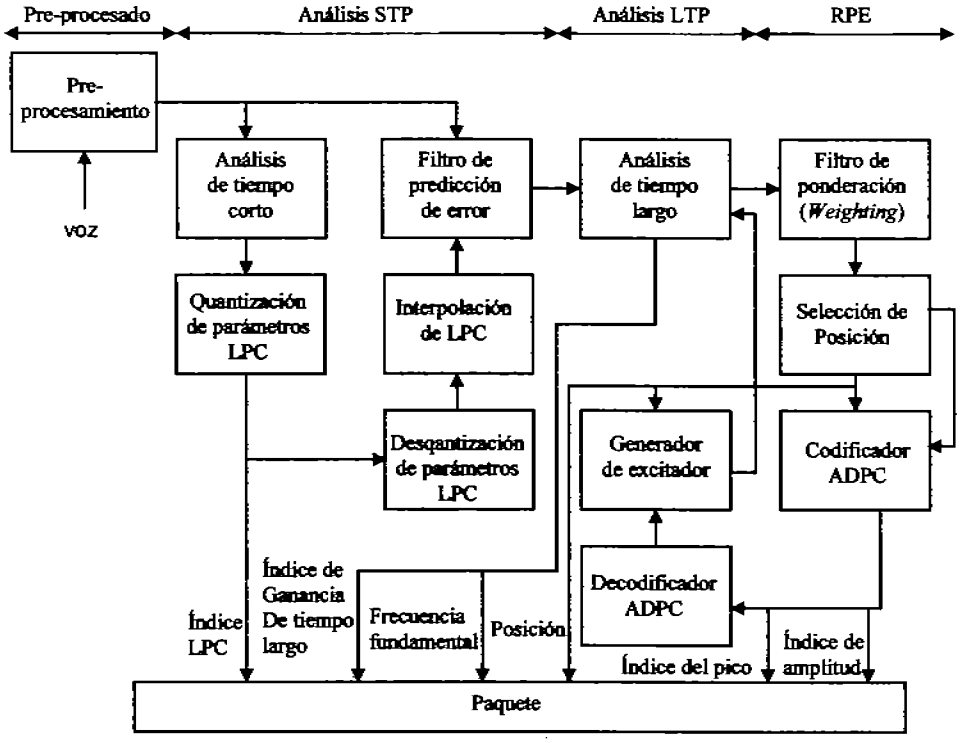
Donde generalmente las funciones recortadoras pueden tener algún tipo de combinación, donde se ha demostrado que [27]:

- Para hablantes con frecuencia fundamental alta, la diferencia en el desempeño de utilizar combinaciones de funciones clipping es pequeña y probablemente insignificante.
- Para hablantes con frecuencia fundamental baja, si existe una diferencia significante en el desempeño con base en los valores de las pruebas, tomando en cuenta que las peores combinaciones son: $s(n)$ con $s(n)$, $s(n)$ con $clc[s(n)]$, $s(n)$ con $clp[s(n)]$ y $s(n)$ con $sgc[s(n)]$.

- El desempeño es mejorado si el procesado no lineal es desarrollado después de un filtro paso bajas.

3.4. Diseños propuestos e implementaciones.

Con base en el diagrama de bloques de la figura 2.1 que a continuación se muestra:



se pueden observar los de diagramas de los cambios generales realizados al estándar descrito en el capitulo pasado para cada uno de sus bloques: bloque de pre-procesado, Análisis STP, Análisis LTP. Así como la simulación tanto del estándar original como las diversas propuestas hechas.

Módulo de Pre-procesado

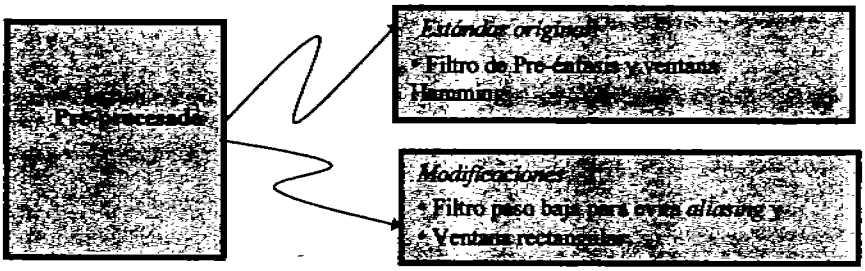


Figura 3.6 Modificaciones Generales en Pre-procesado.

Módulo de Análisis STP.

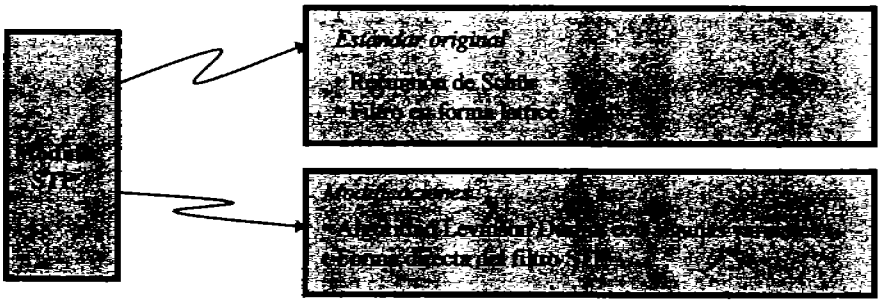


Figura 3.7 Modificaciones Generales en el Módulo STP.

Módulo LTP.

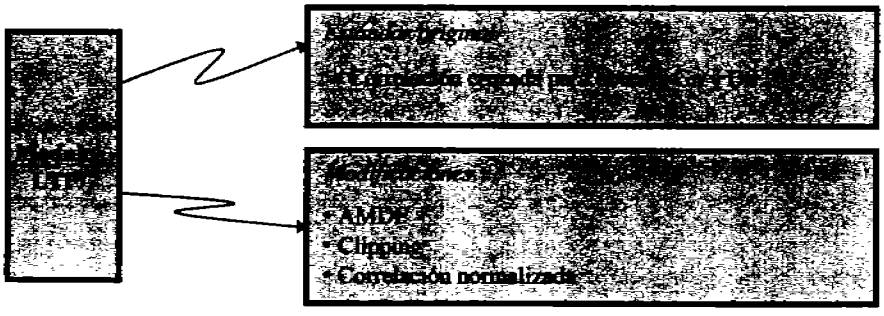


Figura 3.8 Modificaciones Generales en el Módulo LTP.

3.4.1 Simulaciones.- La simulación de los codificadores se desarrollo en Laboratorio de Procesamiento de Voz del edificio de Posgrado de la Facultad de Ingeniería con ayuda del software Matlab versión 6, con una computadora Pentium IV.

A continuación se mencionan las características de dichas implementaciones [28] y se muestran las gráficas de la palabra original y la palabra codificada en tiempo y en frecuencia. Para todas las codificaciones se utiliza la palabra nueve grabada con la voz de una mujer en formato "wav" a una frecuencia de muestreo de 8000 [Hz] y 16 bits mono.

Simulación 1. Estándar original.

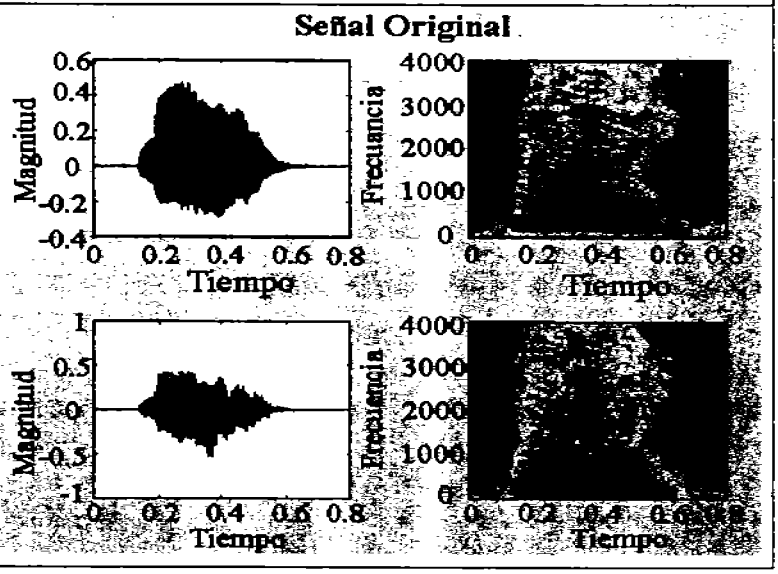
Estándar Original	
Propiedades	***Gráficas en tiempo y frecuencia***
• *Filtro de pre- énfasis* • *Ventana Hamming* • *Recursión de Schür* • *Filtro Lattice* • *Frecuencia fundamental por autocorrelación* • *De- énfasis*	

Figura 3.9 Señal original y señal codificada (inferior) con codificación RPE – LTP y espectograma.

En la figura 3.9 podemos observar la señal original (imagen superior) con su espectro de frecuencia, en la misma figura pero inferior muestra la palabra codificada con el estándar original, podemos ver sus característica donde se utilizo un filtro de pre-énfasis, una ventana Hamming, segmentos de L=160 muestras para obtener 9 coeficientes de autocorrelación de acuerdo a la recursión de Schür y posteriormente ocho coeficientes LAR y la señal residual STP por medio del

Simulación 2. RPT-LTP (Modificación 1).

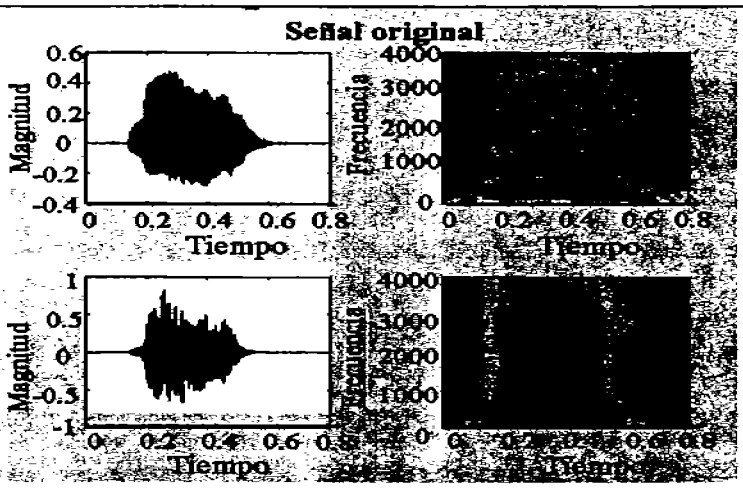
RPE-LTP Modificado 1	
Propiedades	**Gráficas en tiempo y frecuencia**
• Filtro de pre- énfasis • Ventana Hamming • Levinson Durbin • Filtro Lattice • Frecuencia fundamental por autocorrelación • De- énfasis	

Figura 3.10 Señal original y codificada (inferior) con RPE – LTP (Modificación 1) y espectograma.

Simulación 3. RPT-LTP (Modificación 2).

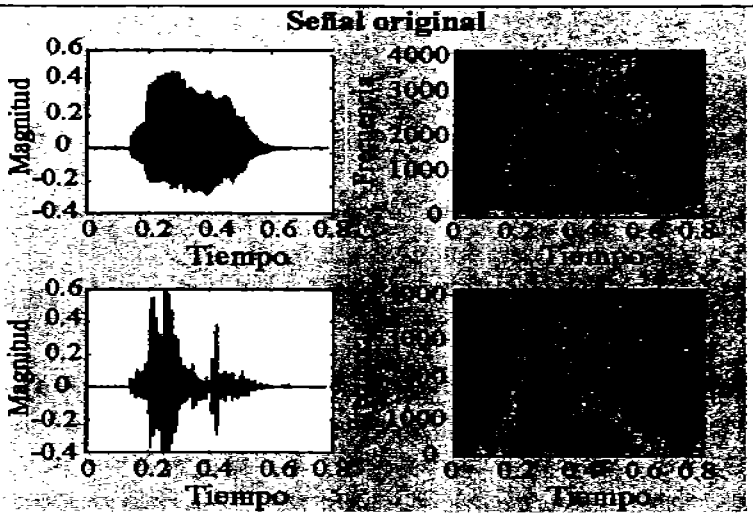
RPE-LTP Modificado 2	
Propiedades	**Gráficas en tiempo y frecuencia**
• Ventana Hamming • Leroux-Gueguen • Filtro Lattice • Frecuencia fundamental por AMDF	

Figura 3.11 Señal original y codificada (inferior) con RPE – LTP (Modificación 2) y espectograma.

Simulación 4. RPT-LTP (Modificación 3).

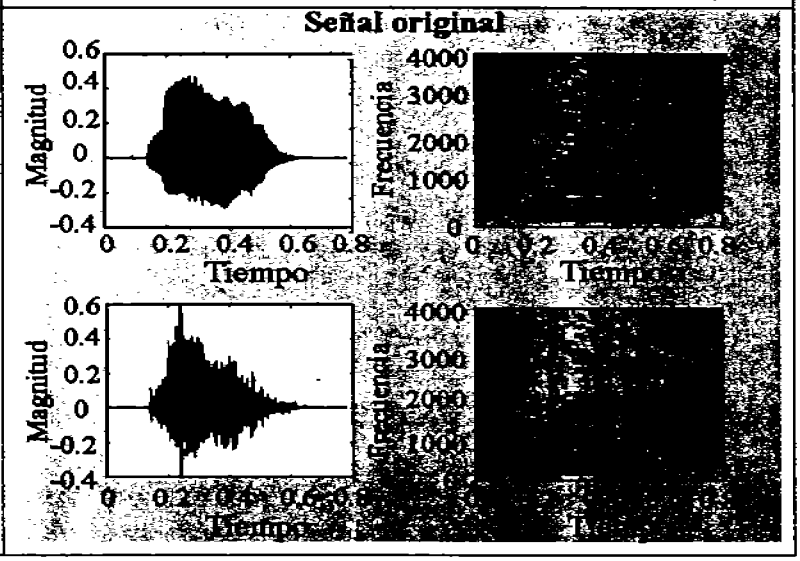
RPE-LTP Modificado 3	
Propiedades	***Gráficas en tiempo y frecuencia***
• *Filtro paso altas* • *Filtro de Pre-énfasis* • *Ventana Rectangular* • *Levinson Durbin* • *Filtro Forma directa* • *Frecuencia fundamental por AMDF con clipping* • *Filtro de De-énfasis*	

Figura 3.12 Señal original y codificada (inferior) con RPE – LTP (Modificación 3) y espectograma.

filtro Lattice, siguiendo subsegmentos de N=40 muestras son necesitados para obtener la periodicidad de la frecuencia fundamental y la ganancia LTP determinadas por autocorrelación y poder aplicar el filtro LTP para por último hacer un codificación y decodificación RPE y un de-énfasis con un coeficiente de a=0.9 igual que en el filtro de pre-énfasis.

En la figura 3.10 el RPE-LTP modificado con algunos cambios propuestos como el uso de algoritmo Levinson Durbin para obtener los coeficientes lpc's y la señal residual STP. Esta implementación es más directa en la obtención de dicha señal residual, sin embargo se convirtieron los coeficientes lpc's en coeficientes de reflexión y a coeficientes LAR para utilizar el filtro Lattice es mostrado con el mismo formato de la figura 3.9.

Otra implementación es mostrada en la figura 3.11 usando el algoritmo *Leroux-Gueguen* y el método AMDF para obtener el retardo D y la ganancia G. Es posible ver que las altas frecuencias se perdieron debido a la falta de uso de los filtros de pre-énfasis y de-énfasis.

Finalmente en la figura 3.12 se presenta una implementación usando un filtro paso altas a la entrada de la señal para eliminar cualquier frecuencia cero (componentes de DC), un filtro de pre-énfasis con un coeficiente de a=0.86, una ventana rectangular, el algoritmo Levinson Durbin, el método AMDF para los parámetros y el filtro de de-énfasis.

Una vez que se logro obtener la simulación de las codificaciones y por lo tanto los archivos *.wav de la voz codificada se procede a medir la calidad de dichas codificaciones, las cuales serán mostradas en el siguientes capítulo.

CAPÍTULO 4
PRUEBAS Y RESULTADO

Para poder probar la calidad de las señales obtenidas en los diseños e implementaciones vistos en el capítulo anterior se diseñaron y realizaron pruebas tanto objetivas, classical signal to noise ratio (SNR), como subjetivas (MOS) a los algoritmos. A continuación se mencionan las características de dichas implementaciones.

4.1 Implementación de pruebas

La implementación de pruebas se llevo a cabo con ayuda de software, con lo que respecta a las medidas objetivas, estas fueron aplicadas sumando al código fuente de los algoritmos simulados las ecuaciones 1.36 y 1.37 vistas en el capítulo 1. A continuación se muestra el código implementado, la tabla y la gráfica de los valores obtenidos para cada una de las simulaciones serán mostrados en los resultados.

Código para SNR.

```
function [s_n_r]=snr(x,xest)
%Función para el cálculo de la relaciuón señal a ruido
%[s_n_r]=snr(x,xest)
%x=señal original
%xest= señal estimada
%s_n_r=SNR en dB
n1=length(x);
n2=length(xest);
if n1<n2
   x=[x zeros(1,n2-n1)];
end
if n1>n2
   xest=[xest zeros(1,n1-n2)];
end
y1=x.^2;
num=sum(y1);
y2=(xest-x).^2;
den=sum(y2);
s_n_r=10*log(num/den)
```

Sin embargo, hay que mencionar que debido a la naturaleza del codificador (híbrido), donde no importa la forma de la onda, sino que suene lo más parecido a la original y debido también a la naturaleza de las pruebas objetivas, las cuales hacen una comparación de que tan parecidas son la forma de ambas señales, las pruebas objetivas no representan un valor real de la calidad de voz en este caso, además de que en ocasiones arrojan valores sin un sentido lógico [23].

4.1.1 Prueba MOS.- La prueba realizada para determinar la calidad de voz fue la prueba MOS, esta fue aplicada en el laboratorio de procesamiento de voz con el inminente ruido de las computadoras y fue respondido por 15 personas, 9 hombres y 6 mujeres.

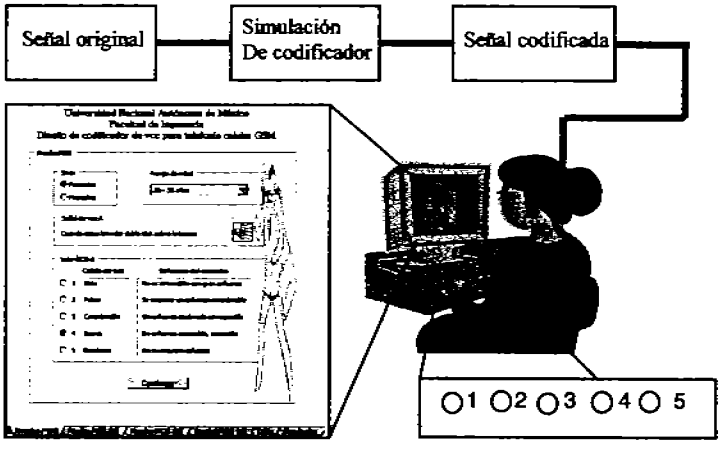


Figura 4.1 Diagrama de prueba MOS.

Los formularios fueros desarrollados en excel donde el participante solo tenía que seleccionar de primera instancia datos personales como sexo y rango de edad con el único fin estadístico, posteriormente con dar un click cuando estuviera listo podría escuchar la señal original, una pausa de 5 segundos y la señal codificada, por último solo se selecciona la opción más indicada en la prueba MOS según su percepción y continuaba (figura 4.2) a escuchar de nuevo la señal original, una pausa de 5 segundos ya la nueva señal codificada y así sucesivamente, hasta que un mensaje mostrando agradecimiento por su participación fuera mostrado.

A continuación se muestran los formularios y el flujo a seguir para cumplir el ciclo de pruebas.

1.

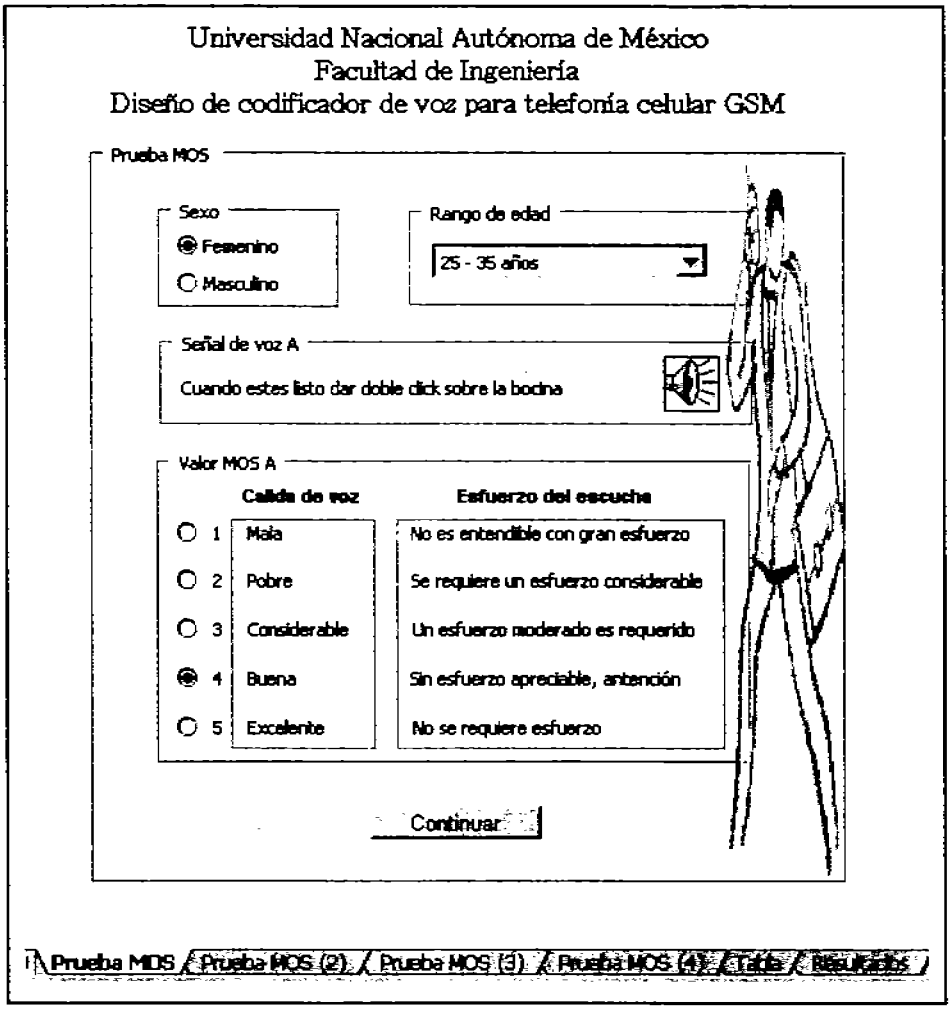


Figura 4.2 Formulario de inicio

En el formulario de la figura 4.2 se escuchará la señal original y la señal codificada con el estándar sin ninguna modificación.

Al dar oprimir el botón de continuar, se mostrara el siguiente formulario

2.

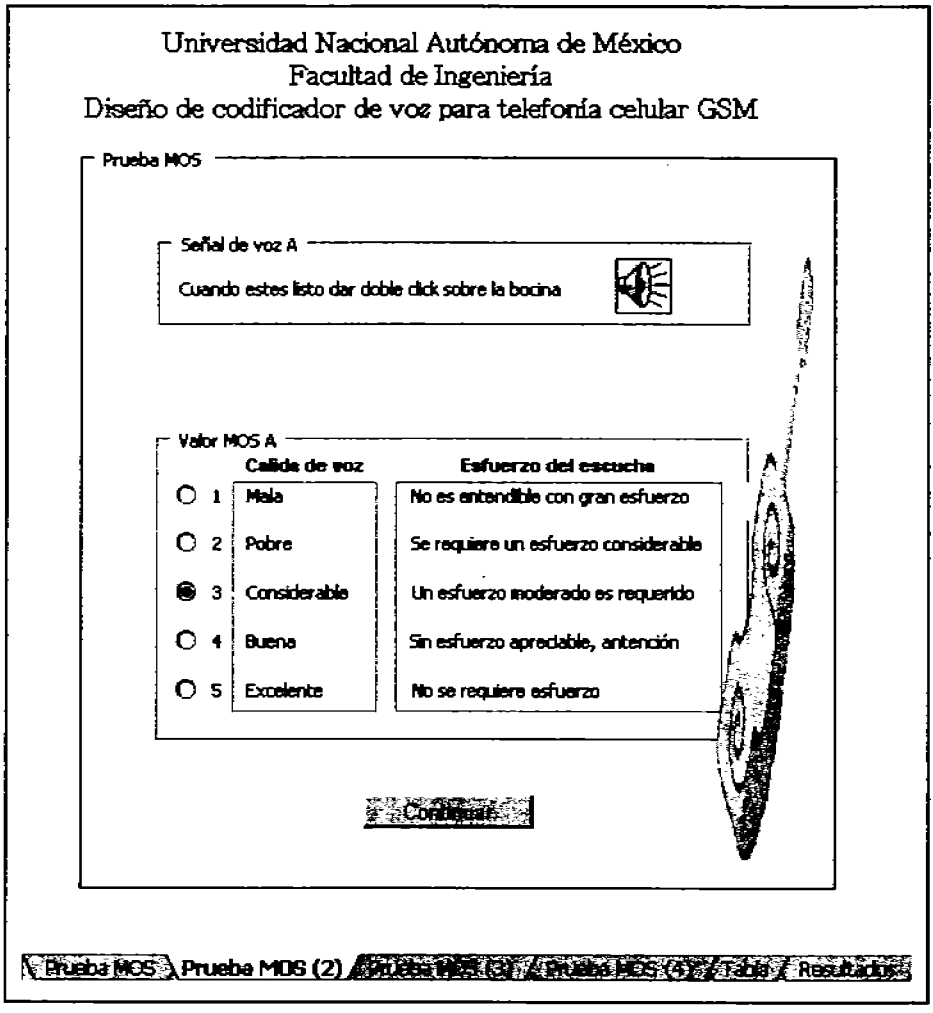


Figura 4.3 Formulario para la codificación modificada 1.

En el formulario de la figura 4.3 se escuchará la señal original y la señal codificada con las características hechas en la modificación 1.

Al dar oprimir el botón de continuar, se mostrara el siguiente formulario

3.

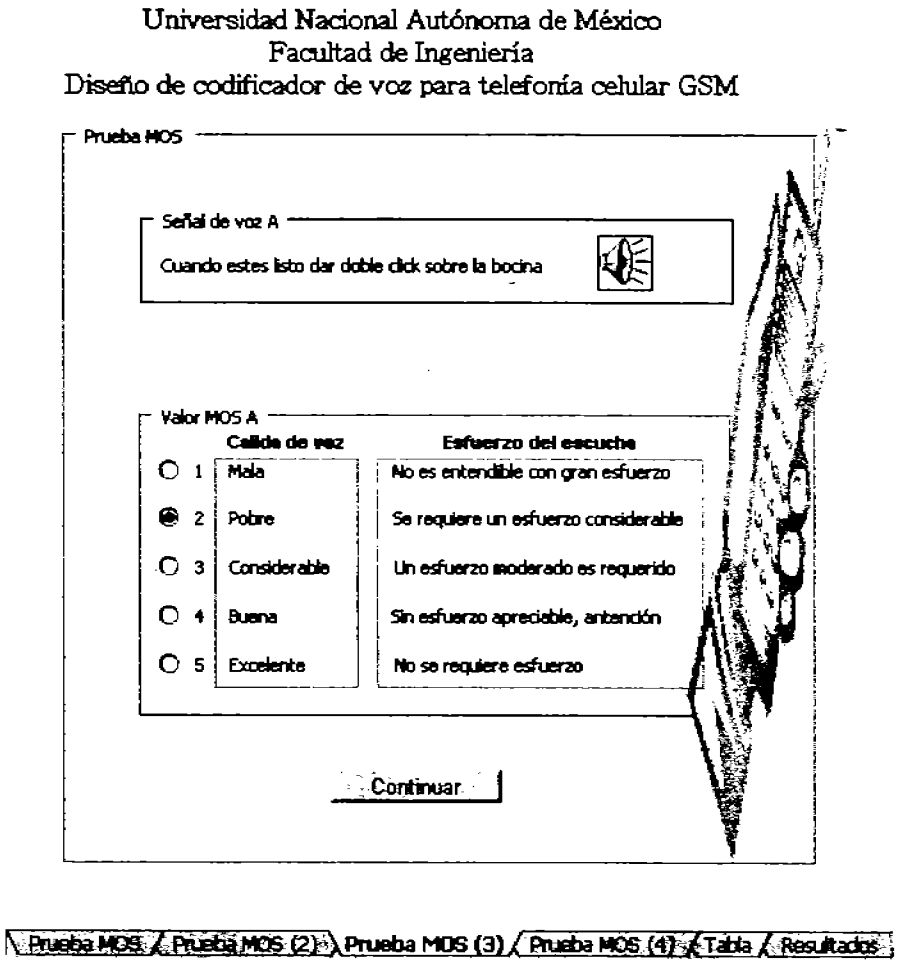


Figura 4.4 Formulario para la codificación modificada 2.

En el formulario de la figura 4.4 se escuchará la señal original y la señal codificada con las características hechas en la modificación 2.

Al dar oprimir el botón de continuar, se mostrara el siguiente formulario

4.

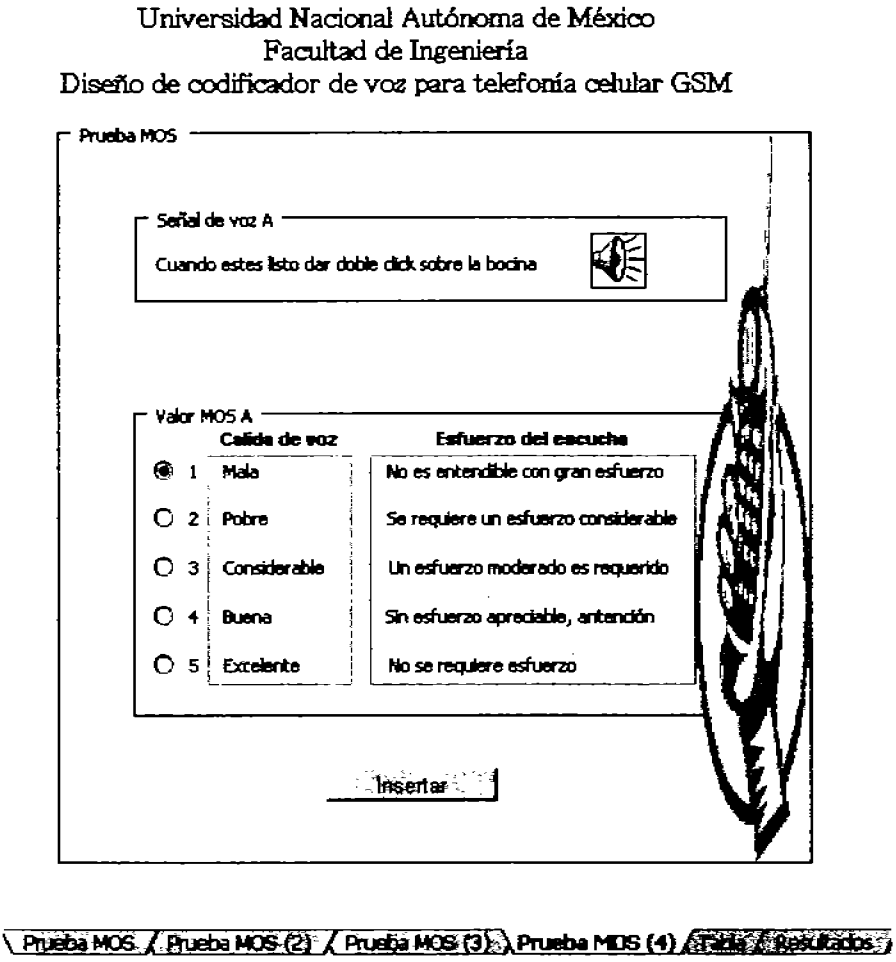


Figura 4.5 Formulario para la codificación modificada 3.

En el formulario de la figura 4.5 se escuchará la señal original y la señal codificada con las características hechas en la modificación 3.

Al dar oprimir el botón de insertar, se acaba la prueba apareciendo el siguiente mensaje.

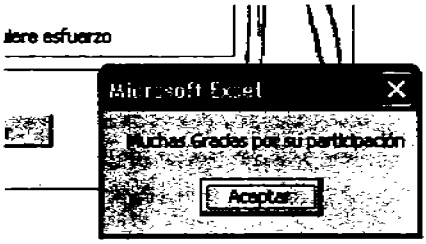


Figura 4.5 Mensaje de agradecimiento.

Los datos insertados por el participante de almacenan en una tabla, mostrada a continuación, por la que podemos acceder a ella por medio de la pestaña Tabla.

			VALOR MOS 2		VALOR MOS 4
2	4	3	4	3	4
2	4	4	3	3	4
2	4	4	5	4	5
1	5	3	4	3	3
2	4	4	3	3	4
2	4	3	4	3	4
1	3	4	4	3	4
1	5	3	4	3	4
2	6	3	4	3	4
1	5	4	4	3	4
1	4	5	3	3	4
2	3	4	4	4	4
2	3	3	4	3	3
1	3	5	5	4	5
2	5	4	4	4	4

Figura 4.6 Tabla de valores MOS introducidos por los participantes

Por último se creo una pestaña para mostrar los resultados, donde se tiene un menú para seleccionar la gráfica que se quiere consultar, teniendo como opción conocer cuantos de los participantes son hombres, cuantos mujeres, cuantos están en determinado rango de edades, ver la gráfica de los valores MOS dado a cada codificación y una gráfica comparativa de todas las codificaciones, como se muestra en la siguiente figura.

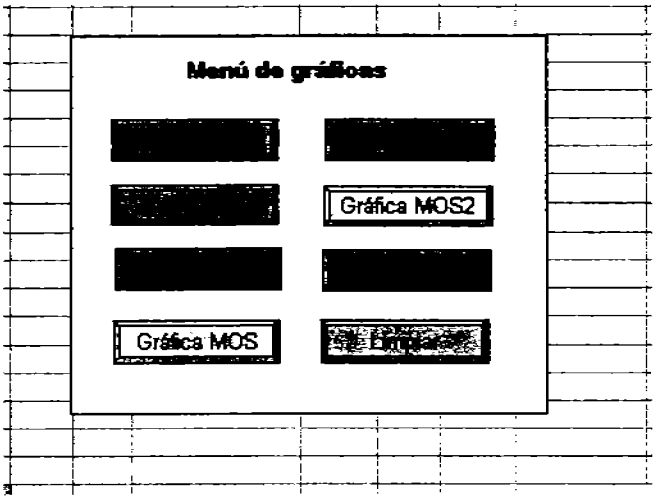


Figura 4.7 Menú de gráficas (Pestaña Resultados).

Los datos que arroja cada menú serán presentados en los resultados de las pruebas.

4.2 Resultados

De esta manera las tablas y gráficas de los resultados son presentadas para tener una mejor idea de cuales de los cambios introducidos y presentados en el diseño del codificador RPT-LTP representan una mejora, sin embargo es importante considerar los siguientes aspectos:

- Al utilizar pruebas subjetiva los resultados se ven alterados con cada escucha, sin embargo al ser los usuarios de dicha codificación quien determine el valor de la calidad, como se había mencionado antes, los resultados son por este tipos de pruebas son válidos en todo sentido.
- El número de escuchas afectará para determinar un valor de MOS adecuado
- La representación de los codificadores en las tablas y gráficas están expresados mediante los nombres asignados en el capítulo 3 es decir:
  - (a) *RPT-LTP (Estándar)*
  - (b) *RPT-LTP (Modificación 1)*
  - (c) *RPT-LTP (Modificación 2)*
  - (d) *RPT-LTP (Modificación 3)*
- Los valores MOS representan el valor de la calidad de la voz en una escala de 1 a 5, considerando el valor MOS=4 de una calidad total.

Medidas objetivas.

CODIFICADOR	SNR [DB]
(a)	-8.7406
(b)	7.6272
(c)	-25.7675
(d)	10.5328

Tabla 4.1 Tabla de valores SNR.

Como podemos observar los valores obtenidos, no representan ningún sentido físico lógico, ya que tenemos valores negativos, lo que supondría que la energía de ruido es mayor a la de la señal. Por lo que como habíamos mencionado, no es una prueba adecuada para este tipo de codificadores.

Medidas subjetivas (Prueba MOS).

Presentamos todas las gráficas y tablas obtenidas por esa prueba.

Femenino	6
Masculino	18

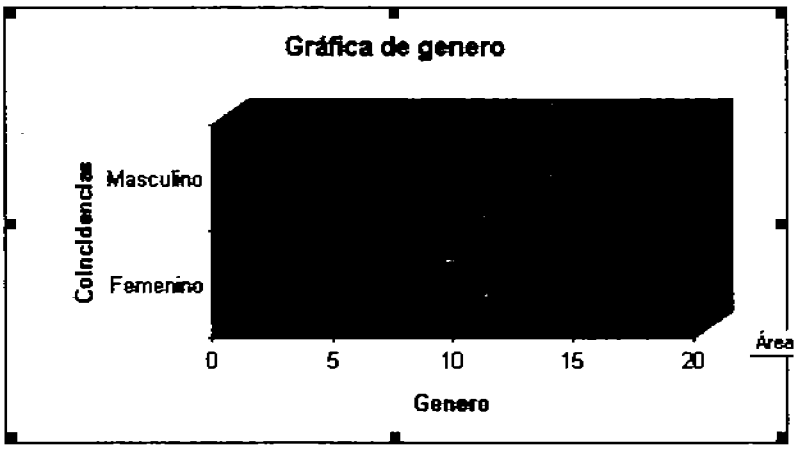


Figura 4.8 Tabla y gráfica de valores referentes al genero de los participantes en las pruebas.

Rango de edades.

Rango	Coincidencia
0-10 años	0
10 - 18 años	4
18 - 25 años	6
25 - 35 años	4
35 - 60 años	1
mayor a 60 años	0

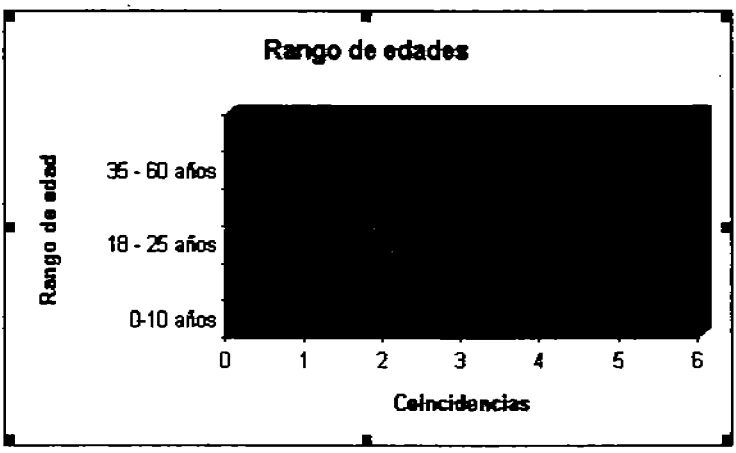


Figura 4.9 Tabla y gráfica de valores referentes al rango de edades de los participantes en las pruebas.

Prueba MOS (a).

Valor MOS	Número de coincidencias
5	2
4	7
3	6
2	0
1	0

MOS:	***3.73333333***

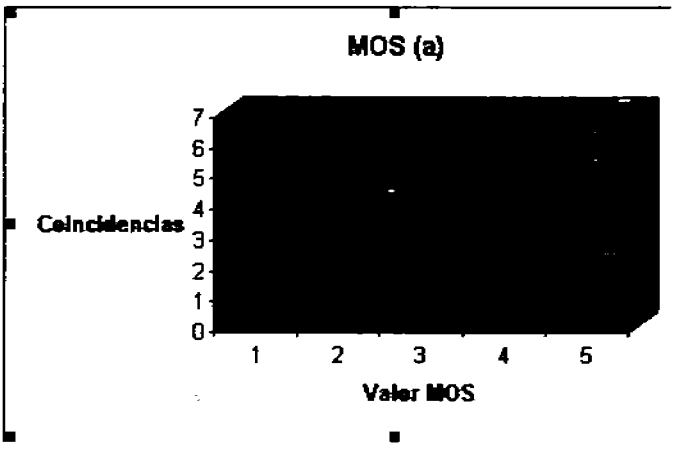


Figura 4.10 Tabla y gráfica de valores MOS asignaos por los participantes para la codificación (a).

Prueba MOS (b).

Valores MOS (b)	
Valor MOS	Número de coincidencias
5	2
4	10
3	3
2	0
1	0

MOS:	***3.93333333***

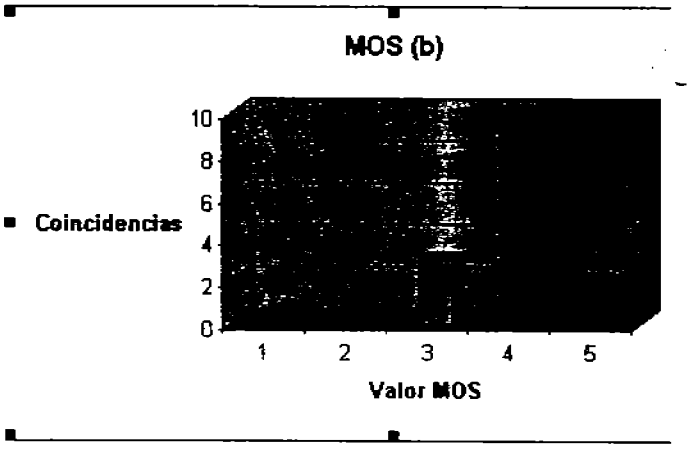


Figura 4.11 Tabla y gráfica de valores MOS asignaos por los participantes para la codificación (b).

Prueba MOS (c).

Valor MOS	Número de coincidencias
5	0
4	4
3	11
2	0
1	0

MOS:	***3.26666667***

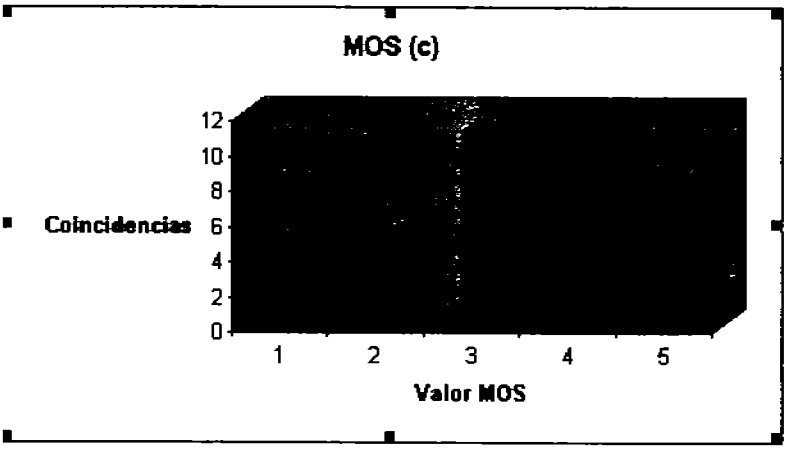


Figura 4.12 Tabla y gráfica de valores MOS asignaos por los participantes para la codificación (c).

Prueba MOS (d).

Valor MOS	Número de coincidencias
5	2
4	11
3	2
2	0
1	0

MOS:	***4***

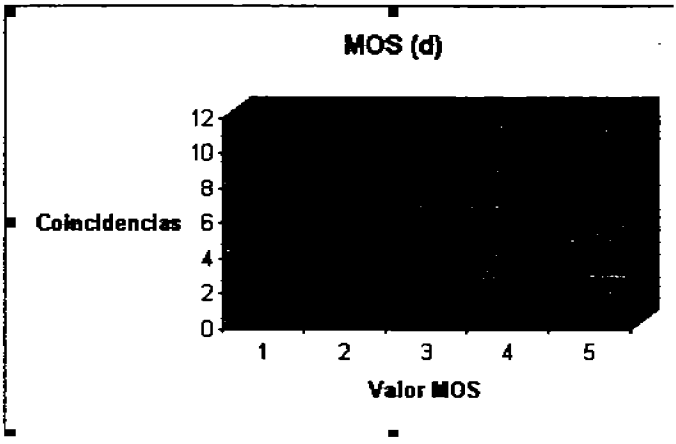


Figura 4.13 Tabla y gráfica de valores MOS asignaos por los participantes para la codificación (d).

Por último se presenta la gráfica del resumen de las pruebas MOS, comparando todas las codificaciones.

Figura 4.13 Tabla y gráfica de valores MOS asignaos por los participantes para la codificación (d).

Conclusiones

De acuerdo con los resultados presentados en el desarrollo de esta tesis y del análisis que se realizó, es posible concluir que la combinación de los diferentes métodos utilizados en toda clase de codificadores de voz en el estándar de telefonía celular GSM RPE-LTP y que por lo tanto al ser utilizados con anterioridad en estándares se puede comprobar su aceptación y eficacia, nos permitió crear un diseño que mejora la calidad de la voz codificada sin alterar en ningún momento características específicas que debe de cumplir un algoritmo de codificación para telefonía celular GSM siendo los más importantes, la velocidad de transmisión en bits por segundo (13 kbps), su retardo y complejidad, al no cambiar la estructura del codificador.

Se desarrollaron tres diseños de codificadores con base en la estructura del codificador RPT-LTP, es decir: una bloque de pre-procesado, análisis STP, una análisis de tiempo largo y post-procesado siendo simulados en Matlab tanto el codificador original como los tres diseños antes mencionados, pudiendo concluir con base en dicho trabajo experimental lo siguiente.

Específicamente para el primer diseño en el cual se hicieron modificaciones solo del bloque de análisis STP al intercambiar el método de la Recursión de Schür para obtención de los coeficientes PARCOR por el método de Levinson Durbin se obtuvo una mejoría en la calidad de voz subiendo de un valor MOS de 3.7 a 3.9 pudiendo apreciar ligeramente la mejoría al escuchar las dos señales codificadas, pero no teniendo un contraste significativo.

En el segundo diseño, en el cual se elimina casi por completo los módulos de pre-procesado y se cambian los métodos tanto en el análisis STP como en el de LTP, pudimos observar que la calidad disminuyo de manera considerable al compararla con la señal codificada estándar pasando de un valor MOS de 3.7 a 3.3, así en este caso el principal contraste es la perdida de las frecuencias altas.

Con el último diseño y con base en las experiencias anteriores, se dio énfasis a los bloques de pre-procesado incorporando un filtro paso altas para eliminar los componentes de DC, además de incorporar el método Levinson Durbin el cual ya no había dado una respuesta positiva. Con lo que respecta al módulo de LTP, se incorporo el método AMDF haciendo el proceso más rápido y con menos carga computacional al sustituir las multiplicaciones por sumas además de utilizar una función clipping, concluyendo con base en las pruebas experimentales, que este diseño arrojó la mayor calidad en la señal codificada obteniendo lo que se considera una calidad total, es decir un valor MOS=4 contra el 3.7 de la codificación original.

Con lo que también podemos concluir que el principal factor que puede llegar a influir en calidad de voz son los procesos para la obtención de la frecuencia fundamental, con lo que si se tienen procesos mucho más exactos y sensibles en la obtención de dicho factor la calidad de voz codificada alcanzaría la calidad total y perdería su sonido artificial.

Por último con lo que respecta a las pruebas para medir la calidad del proceso experimental se implementaron las dos vertientes en estos procesos: las pruebas objetivas, en las cuales deberíamos de tener la certeza de que el valor arrojado por este tipo de pruebas marca la calidad de la señal, sin embargo, fue posible comprobar que no es así debido a la naturaleza tanto de las pruebas, como en la naturaleza del codificador. En este caso en particular, estamos hablando de la relación señal a ruido SNR, la cual compara muestra a muestra, por medio de la energía, la similitud de dos señales, pero cuando este método es probado para una codificación de tipo híbrida, como es el caso de RPT-LTP, donde en realidad no importa que las señal original sea lo más parecida a la codificada, sino que la señal codificada suene lo más parecido a la señal original, este puede arrojar valores que no corresponden a una calidad del codificador como lo vimos incluso al tener vales negativos.

Por lo que con base en lo anterior también pudimos comprobar que la segunda vertiente, la de las pruebas subjetivas es la mejor opción para este tipo de codificación, por lo que se optó en implementar la prueba MOS, con la cual se desarrollo una serie de formularios en Excel, pudiendo programar todo el proceso haciendo más rápido y menos costoso, teniendo los resultados por lo menos parciales de una manera automática.

BIBLIOGRAFÍA

[1] Juang, B.H., The Past, Present, and Future of Speech Processing, *IEEE Signal Processing Magazine*, May 1998, 24-48

[2] Zhong, Yi X. Advances in Coding and Compression, *IEEE Communications Magazine*, July 1993, 70-72, 353-363

[3] Chu, W.C., *Speech coding algorithms* (John Wiley & Sons, Inc., 2003).

[4] Herrera, Abel, *Apuntes de Procesamiento Digital de Voz* (UNAM, 1999).

[5] Flores, Andres, Reconocimiento de palabras aisladas. Pontificia Universidad Católica del Perú. *http://www.alek.pucp.edu.pe/~dflores/tesis/modelo.html*. 1998.

[6] Deller, Proakis y Hansen, *Discrete time processing of speech signals*, (New Jersey: Prentice Hall, 1987).

[7] Rabiner L., *Fundamentals of speech recognition* (New Jersey: Prentice Hall, 1993).

[8] Papamichalis, Panos E., *Practical approaches to speech coding*, (New Jersey: Prentice Hall, INC, 1987)

[9] Medan, Yoav, Yair Eyal y Chazan Dan, Super resolution Pricth Determination of Speech Signals, *IEEE Transaction on Signal Processing*, vol. 39 No. 1, January 1991

[10] Kroon, Peter y Deprettere, Ed F., A Class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbits/s, *IEEE Journal on Selected Areas in Communications, 6(2)*, 1988, 353-363

[11] González, Joaquín, Panorámica de los esquemas de mejora de voz en presencia de ruido, *SEAF I Congreso de la Sociedad Española de Acústica Forense*, 25-41, octubre de 2000.

[12] Natvig E. Jon, Evaluation of Six Medium Bit-Rate Coders for the Pan-European Digital Mobile Radio System, *IEEE Journal on Selected Areas in Communications, 6*(2), 1988, 324-331

[13] Goodman J., David, An IEEE Meeting on Subjective Testing of Voiceband, *IEEE Communications Society Magazine*, 4-5, November 1977.

[14] Kitawaki, Nobohiko, Quality Assesment of Coded Speech, *IEEE Transaction on Signal Processing*

[15] Llisterri, J. y Poch, D., Caracterización fonética del bilingüismo, análisis acústico del habla espontánea y evaluación de sistemas de síntesis de habla, *Simposio de la Lengua Española. Ciencia y Tecnología*, Barcelona, Octubre 1991

[16] Gerson, I.A y Jasiuk, M.A., Vector sum excited linear prediction (VSELP) speech coding at 8 kbps. *Proceedings, IEEE international Conference Acustics, Speech and Signal Processing*,461-464, April, 1990

[17] Lajos, Hanzo, *Voice compression and communications: Principles and applications for fixed and wireless channels*, (USA: Wiley-Interscience, 2001).

[18] Mouly, Michel, Current Evolution of the GSM Systems, *IEEE Personal Communications*, October 1995

[19] Proakis , J. y Manolakis, D., *Digital Signal Processing: Principles, Algorithms and Applications*, (New Jersey: Prentice Hall, 1995) pp. 868-873.

[20] Wade, Graham, *Coding techniques: an introduction to compression and error control*, (Great Britain, 2000)

[21] Bekkers, Rudi, *Mobile telecomunications: Standards, Regulations and Applications*, (Boston: Artech House Publishers, 1999. P.188)

[22] http://mia.ece.uic.edu/~papers/WWW/MultimediaStandards/chapter3.pdf

[23] Valdés, Judith, *Análisis de codificaciones de voz utilizadas en telefonía celular GSM*, Tesis de licenciatura-UNAM, (México, Marzo de 2004)

[24] Le Roux, J. y Gueguen, C. A fixed point computation of partial correlation coefficients, *IEEE Trans. on Acoustics, Speech and Signal Processing 25:3*, Jun 1977, 257-266.

[25] Hess, W. Pitch Determination of Speech Signal-Algorithm and Devices. *Springer Series in Information Sciences*, (Germany, 1983)

[26] Rabiner , L., Cheng J. y Rosenberg A.E. A comparative performance study of several pitch detection algorithms. *IEEE Trans. on ASSP, 24(5)*, October 1976, 399-418.

[27] Rabiner, L., On the use of autocorrelation analysis for pitch detection. *IEEE Trans. on ASSP*, February 1977, 24-33.

[28] Valdés, Judith y Herrera, Abel, Improvements for the RPE-LTP speech codec, *The Seventh IASTED International Conference on Signal and Image Processing*, 23-27, Honolulu, August 2005.