Estudio de Sistemas de compresión de voz digital orientado a telefonía celular.

Javier Alejandro Bustos Jiménez

1 Introducción

Últimamente las empresas de telefonía celular han tenido problemas para satisfacer la gran demanda de comunicación que se ha originado. En Chile, la banda en la que pueden transmitir dichas empresas es licitada por la subsecretaría de telecomunicaciones (Subtel). Estas bandas de transmisión son excluyentes y por lo tanto un recurso escaso. Además, se espera una penetración (abonados por cada 100 habitantes) de un 40% dentro de 7 a 10 años (en 1999 era de 15%) [8].

En condiciones normales de transmisión de voz, se utiliza una tasa de muestreo de 8.000 datos por segundo [4], codificándola en 8 bits, por lo que se necesita 64 Kbps de capacidad de transmisión. Con los métodos de compresión utilizados actualmente, se necesitan sólo 8-13 Kbps de capacidad y la señal está catalogada como buena.

En el año 1996 un alumno de la FCFM de la Universidad de Chile desarrolló un sistema compresor digital de voz basado en redes neuronales artificiales (RNA) [1] del tipo Mapas Autoorganizativos y Backpropagation [2][3][5], el cual necesitaba de 6,4 Kbps de capacidad y generaba una calidad de señal medianamente buena (según test MOS). Pero, se vio imposibilitado de realizar un estudio comparativo con los algoritmos utilizados en telefonía celular puesto que no se contaban con los medios para reunir esa información y realizar la experimentación.

Este trabajo tiene por finalidad el facilitar la realización de este tipo de estudios, para ello se reunió información teórica (algoritmos) y práctica (códigos fuente) de los tres sistemas de compresión más utilizados en la actualidad: CELP (creado para la comunicación radial federal de Estados Unidos y utilizado actualmente como estándar de compresión de voz de la telefonía celular digital); VSELP (creado a partir de CELP por la empresa MOTOROLA y estandarizado para su uso en telefonía celular análoga) y GSM 06.10 (estándar de telefonía celular utilizado en Europa); y de una versión mejorada del algoritmo basado en RNA [11]. Se realiza además un estudio comparativo de estos 4 algoritmos.

2 Herramientas aplicadas al análisis de la voz

A continuación se presenta el marco teórico en los cuales están basados los sistemas de compresión de voz de este estudio. En primer lugar se presenta el funcionamiento del aparato fonador humano y sus características físicas. Luego se presenta la teoría de la señal y las herramientas para su representación matemática. En tercer lugar se presenta el concepto de filtro para llegar al modelo fuente/filtro y por último se presenta la teoría de predicción lineal.

2.1 Modelamiento del tracto vocálico humano

"La voz se produce a partir de sonidos formados por la vibración de las cuerdas vocales y posterior resonancia en la pared del tracto vocálico de la señal producida. En los adultos, el tracto vocálico es un tubo de aproximadamente 17 cm de largo con un área transversal que varía de 0 a 20 cm2 "[4]. La figura 1 muestra un diagrama del tracto vocálico.

Los pulmones actúan solamente como emisores de aire. Son las cuerdas vocales las encargadas de introducir una perturbación cuasi periódica en el flujo de aire.

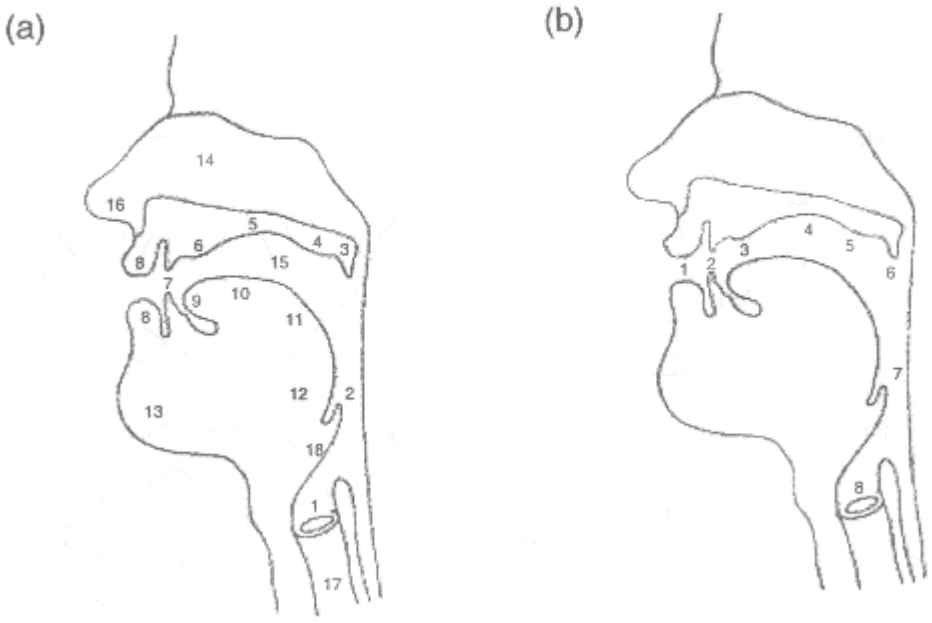


Figura 1 [10]: Tracto vocálico. a) articulaciones del habla: (1) cuerdas vocales; (2) faringe; (3) velo; (4) paladar blando; (5) paladar duro; (6) alveolos; (7) dientes; (8) labios; (9) punta de la lengua; (10) cuerpo lingual; (11) dorso; (12) raíz; (13) mandíbula; (14) cavidad nasal; (15) cavidad oral; (16) ventanas nasales; (17) traquea; (18) epíglotis. b) tipos de articulación de voz: (1) labial; (2) dental; (3) alveolar; (4) palatal; (5) velar; (6) uvular; (7) faringeal; (8) glotal.

Los sonidos que conforman la voz se pueden clasificar en vocalizados (sonoros, originados en las cuerdas vocales) y no vocalizados (sordos, originados por una fricción en el tracto vocálico), en la práctica la voz está formada por una mezcla de ambos.

Durante el proceso de generación de sonidos vocalizados, las cuerdas vocales están cerradas, pero la presión ejercida por el aire contenido en los pulmones fuerza su apertura y su posterior relajación ocasionando la vibración de las cuerdas a una frecuencia entre los 50 y 400 [Hz]. A esta frecuencia se le conoce como pitch.

La forma de la señal que se produce en la vibración con las cuerdas vocales es aproximadamente triangular. Ésta atraviesa el resto del tracto vocálico donde la amplitud se ve alterada por el choque de la señal con las paredes del tracto.

Durante el proceso de generación de sonidos no vocalizados, las cuerdas vocales están completamente abiertas, posibilitando la circulación del aire por el tracto vocálico, la que se ve ligeramente obstaculizada por el roce con las paredes del tracto, lo que produce un ruido fricativo.

Además del movimiento de las cuerdas vocales y del tracto vocálico, para modelar el proceso de generación de voz se debe considerar también los movimientos de la boca, la lengua, los labios y vibraciones nasales. Por tanto, un modelo básico de este proceso debe considerar lo siguiente:

- La voz es una señal que emerge de una fuente definida: los pulmones actúan como emisores de aire y la señal se produce por la vibración de las cuerdas vocales y la posterior resonancia con las paredes del tracto vocálico.
- La voz está formada por la mezcla de señales de excitación periódica y ruido.
- La variación temporal de la señal en el tracto vocálico produce el timbre característico que diferencia los fonemas, ciertos fonemas son articulados sin la presencia de las cuerdas

vocales (fonemas sordos).
- Antes de pasar por el tracto vocálico, la onda sonora tiene un espectro relativamente plano (sin formantes).
- La fuente emisora posee dos estados: generación de sonidos vocalizados y no vocalizados.
- Si se toman intervalos de tiempos pequeños se puede modelar el órgano generador de voz a través de la búsqueda de su función de transferencia, que define relación entre la entrada (excitación glótica) y la salida (voz generada) por medio de filtros.

Los sistemas de codificación de voz que utilizan éste modelo se denominan VOCODER y se utilizará esa nomenclatura a lo largo de este documento.

2.2 Filtros Lineales

La producción de voz es modelable por sistemas matemáticos conocidos como filtros lineales, los cuales se utilizan para representar la función de transferencia del sistema H(z): la voz puede ser considerada como una señal que se genera a partir del filtrado de la excitación glótica más una señal de amplitud aleatoria de espectro uniforme, conocida como ruido blanco (Figura 2).

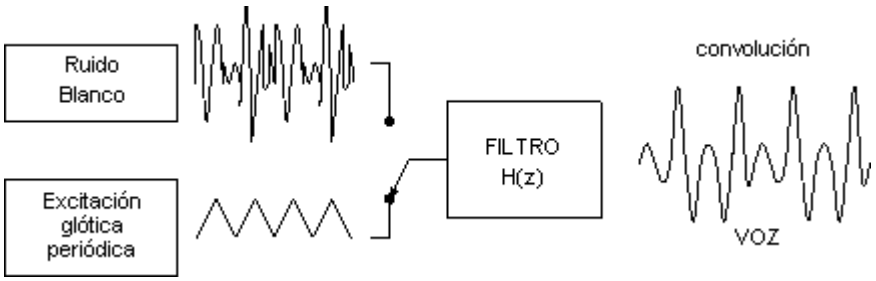


Figura 2: La voz se produce del filtrado de la excitación glótica y el ruido blanco.

La estimación de parámetros lineales es una de las técnicas más utilizadas para encontrar la función de transferencia de un filtro y que caracteriza al sistema. Su operación se basa en que la evolución de un sistema depende de entradas particulares y respuestas pasadas del sistema.

2.3 Estimación de parámetros de un filtro lineal

Para el análisis (y síntesis) de la voz se utiliza el modelo fuente/filtro [4], basado en una batería de filtros lineales y causales que modelan los distintos componentes del aparato fonador humano. Para la estimación de los parámetros de este filtro se utiliza la predicción lineal denominada LPC que se estudia en la sección 2.3.1.

2.3.1 Predicción lineal (LPC)

El análisis de LPC se utiliza para encontrar los coeficientes que representarán la función de transferencia del filtro que modela el sistema. Si el modelo es capaz de predecir la señal con un error muy bajo, se tiene que el LPC ha sido capaz de almacenar la información necesaria de un trozo de señal como para reproducirla mediante alguna excitación. En analogía con un instrumento musical, el LPC sería un instrumento de viento que al ser soplado emite el sonido con el timbre particular del trozo de voz que representa.

El principio de un LPC es que el valor actual de una muestra de señal de voz, s(n), puede predecirse a partir de un número finito de muestras anteriores: s(n-1), ... , s(n-p), con un error asociado e(n) utilizando un filtro lineal sólo polos:

p

$$s(n) = e(n) + \sum_{k=1} \alpha_k \, s(n - k)$$

El error de predicción (también conocido como señal residual), e(n), es simplemente la diferencia entre el valor actual de la señal, s(n), y el valor que se predijo, ŝ(n):

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

Los factores que otorgan el peso, α_k, son encontrados al minimizar el error cuadrático medio, encontrado en N muestras (E):

$$E = \left(\sum_{i=0}^{N-1} e^2 (i)\right) / 2$$

Los coeficientes α_k que minimizan el error de predicción E_n son calculados igualando el gradiente con respecto a α_i a 0, para i = 1, .., k. Lo que da como resultado una serie de ecuaciones lineales:

$$(\partial E/\partial \alpha_k) = 0 \quad \forall k$$

Se sabe que s(n) es constante, porque es la señal original, luego al derivar para encontrar el mínimo se tiene que:

$$(\partial E/\partial \alpha_k) = \sum_i (s(i) - \sum_j \alpha_j \, s(i - j)) \, s(i - k) = 0. \quad j = 1 \,..\, p$$

$$\sum_i (s(i) \cdot s(i - k)) = \sum_j \alpha_j \left(\sum_i s(i - j) \, s(i - k)\right)$$

Si se define $\gamma(i,k) = \Sigma_n \, s(n - i) \cdot s(n - k)$ se obtiene un sistema de ecuaciones matricial de la forma:

$$\begin{matrix} \gamma(1,1) & \cdots & \gamma(1,p) \\ \cdots & \cdots & \cdots \\ \gamma(p,1) & \cdots & \gamma(p,p) \end{matrix} \begin{matrix} \alpha_1 \\ \cdots \\ \alpha_p \end{matrix} = \begin{matrix} \gamma(1,0) \\ \cdots \\ \gamma(p,0) \end{matrix}$$

Pero se tiene un problema, para medir en los bordes de la ventana de la señal se necesita salir de la ventana, por lo tanto se pueden hacer 2 suposiciones:

1. Medir fuera de la ventana suponiendo que es cíclica.
2. Todo fuera de la ventana es 0, lo cual otorga una nueva función g cuyos límites son -∞ y +∞, esto simplifica bastante el problema, llegando al siguiente estado:

$$\gamma(i,k) = r_{|i-k|} = \sum_{n=-\infty}^{+\infty} s(n) \, s(n - (i-k))$$

La ecuación anterior bajo esas condiciones es conocida como método de autocorrelación. Esto es casi equivalente a suponer que la señal se repite y que la función fuera de la ventana vale 0,

por lo tanto la matriz adopta la siguiente forma.

$$\begin{matrix} r_0 & \cdots & r_{p-1} \\ \cdots & & \cdots \\ r_{p-1} & \cdots & r_0 \end{matrix} \begin{matrix} \alpha_1 \\ \cdots \\ \alpha_p \end{matrix} = \begin{matrix} r_1 \\ \cdots \\ r_p \end{matrix}$$

Además, $r_k = \Sigma_n\ s(n)\ s(n- k)$ con n = k, ..., N-1 y k = 0, ..., p. Es decir, basta calcular p+1 valores en la ventana, esquemáticamente se presenta el analizador LPC en la figura 3.

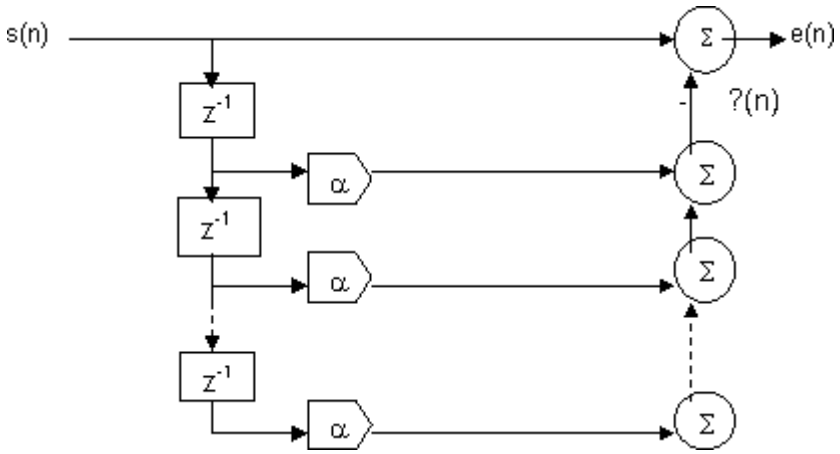


Figura 3: Esquema de un analizador LPC

3 Algoritmos de Compresión de Voz

En este capitulo se estudiarán cuatro algoritmos de compresión utilizados para realizar el estudio comparativo.

3.1 CELP

Todas las técnicas de compresión de voz están basadas en dos operaciones intrínsecas:

- Eliminar la redundancia.
- Eliminar la irrelevancia.

La primera operación utiliza predicciones o transformaciones para eliminar los datos redundantes, lo cual reduce el ancho de banda necesario para la señal. La segunda operación reduce el ancho de banda realizando una cuantización de, ya sea los componentes de la predicción (o su error) o de los coeficientes de la transformación. Obteniendo una señal parecida a la original pero siempre con un grado de distorsión o error de reconstrucción.

Al aumentar la compresión, es necesario que el codificador minimice la percepción del error utilizando propiedades inherentes al habla humano. Esto quiere decir que el mismo nivel de error de la distorsión es percibido de distinta manera si es aplicado a señales de voz con distinta energía y bandas de frecuencia.

La solución de CELP a ese problema es utilizar la aproximación análisis por síntesis, donde se mide la percepción de la distorsión.

Un codebook consiste en una tabla de muestras de señal residual, conocidas como codewords, los cuales se utilizarán como excitación de los filtros. Además, un filtro llamado "de peso de percepción", es utilizado para asegurar que la medida del error cuadrático medio refleje el error de percepción.

Al aplicar un filtro de percepción sobre la señal se mejora el rendimiento del codificador. Los formantes de alta energía disimulan mejor el ruido que las porciones de baja energía del espectro. La señal de error generada por cada paso del sintetizador es ponderada apropiadamente para mejorar este efecto de percepción. El filtro amplifica la señal de error en las regiones en que no hay formantes y lo atenúa en las que sí. De este modo, una señal de error cuya energía es concentrada en los formantes es considerada mejor que una que no.

En la práctica (Figura 4), los sistemas CELP emplean algoritmos rápidos de búsqueda explotando la estructura computacional de éste. Es por eso que el esquema original derivó en un nuevo esquema:

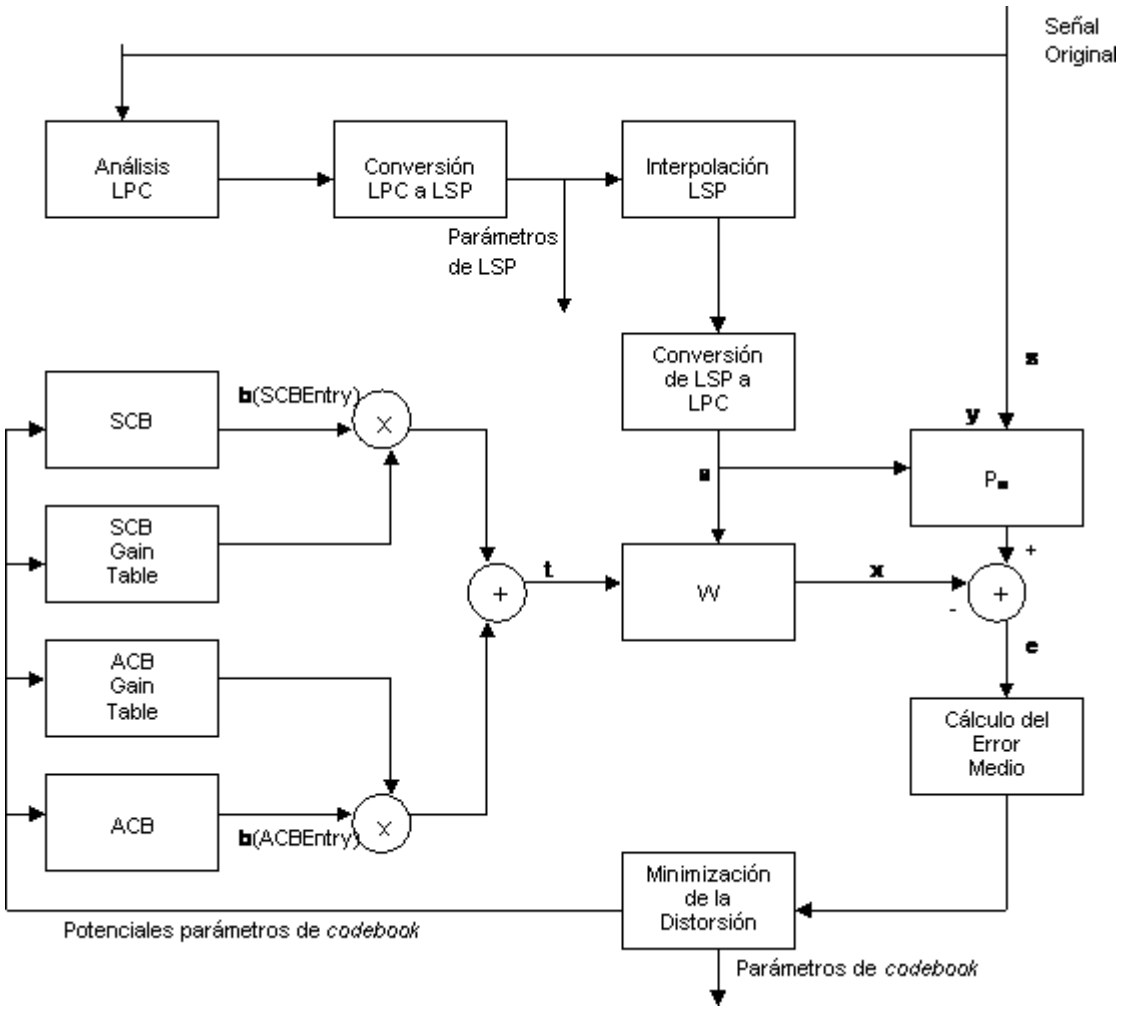


Figura 4: Esquema de un analizador CELP en la práctica

El decodificador toma los parámetros codificados y utilizando el mismo esquema, pero en sentido inverso, reconstruye la señal ŝ. Además, se encarga de sincronizar la señal construida del ACB, para ello, utiliza las dos últimas muestras del subframe anterior.

3.2 VSELP

VSELP fue diseñado por Motorola quien es el responsable del diseño y desarrollo del algoritmo.

VSELP es un tipo de algoritmo CELP que codifica a 7.950 [bps], utilizando un adicional de 5.050 [bps] para control de errores y sincronización de tramas.

La diferencia entre VSELP y los codificadores comunes (CELP) radica principalmente en la estructura de sus codebooks. Mientas CELP utiliza un stochastic codebook para realizar sus

búsquedas, VSELP utiliza dos conjuntos de vectores base para generar un espacio de "vectores candidatos". De este modo, la búsqueda en el codebook de CELP corresponde a dos búsquedas en VSELP.

Hay siete vectores base ortogonales [12] en todo el espacio para cada búsqueda. Cada uno de ellos contiene 40 elementos. La selección de los vectores base es fundamental para una rápida búsqueda en los codebook.

Se realiza un análisis LPC para cada ventana de señal de voz y se obtienen una serie de coeficientes del filtro LPC. Estos coeficientes son expandidos en bandas de frecuencia para utilizarlos en un filtro de error perceptual.

El análisis por síntesis procede con 3 codebook. Primero, se busca en el adaptive codebook, obteniéndose "la mejor" entrada y ganancia. Esta entrada multiplicada por su factor de ganancia es ortogonalizada para el espacio de los primeros 7 vectores base. De este modo, la búsqueda en el segundo codebook puede realizarse independiente del primero.

Un nuevo espacio de 7 vectores es utilizado para la segunda búsqueda; y una nueva "mejor" entrada y ganancia se obtienen de este segundo codebook, ortogonalizandola como en la etapa anterior.

Finalmente se realiza una búsqueda en un tercer codebook. La ganancia obtenida de cada uno de los 3 codebooks se cuantiza y transmite con los 3 índices del codebook al receptor.

Los principales módulos de VSELP (Figura 5) son:

- Análisis de LPC de orden 10.
- Predicción de largo plazo.
- Búsqueda en el adaptive codebook (pitch).
- Búsqueda de la primera base de vectores en el codebook.
- Búsqueda de la segunda base de vectores en el codebook.
- Cuantización vectorial del codebook de la ganancia.

Este algoritmo utiliza una frecuencia de muestreo de 8 [Khz], 160 muestras por frame, dividiéndola en 4 subframes de 40 muestras. El factor de expansión de banda es 0.8.

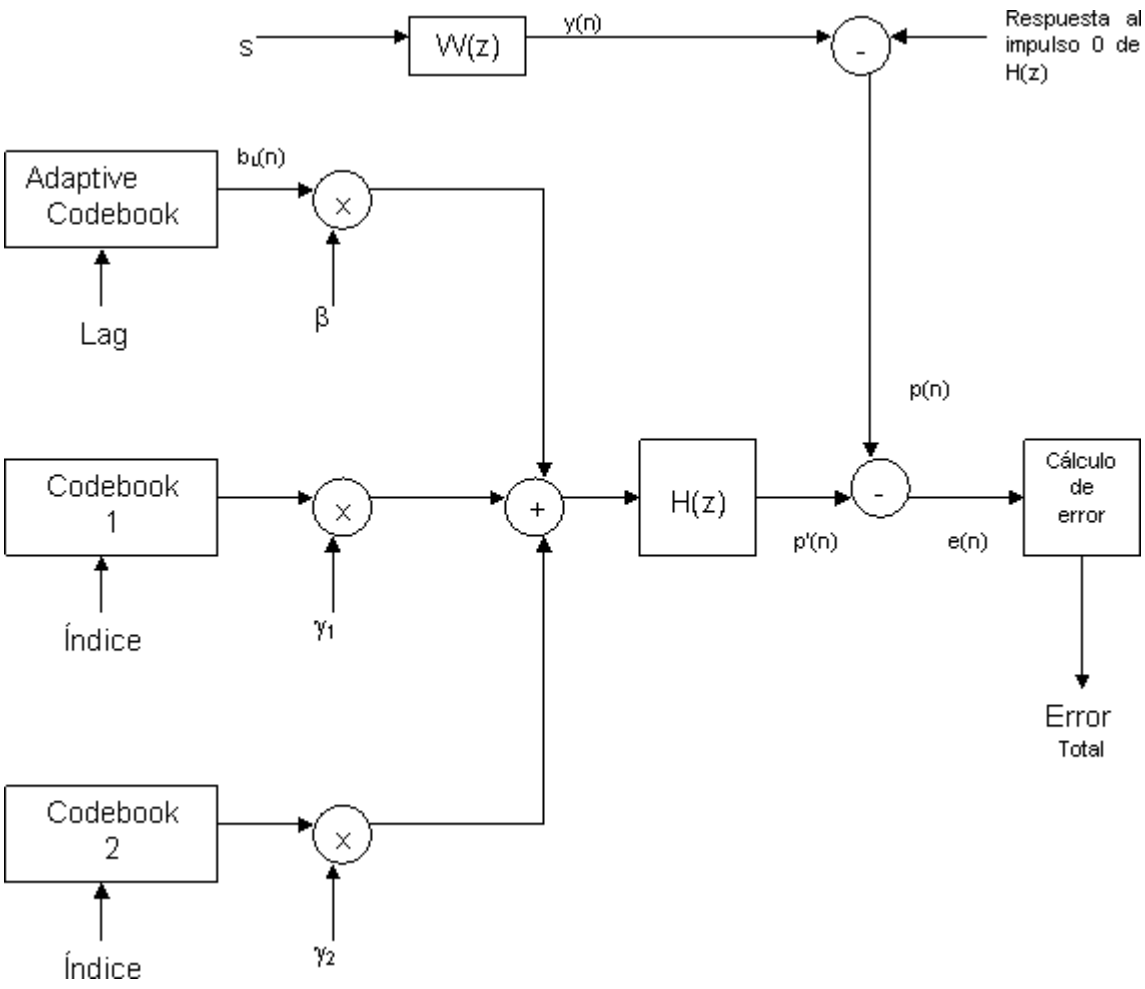


Figura 5: Esquema del analizador VSELP

Si se toma γ_2=0 y $b_L(n)$ como dato para la primera búsqueda en los codebooks, se obtendrán los valores óptimos de β, γ_1 y $f_{1,I}(n)$. Un procedimiento similar se realiza para γ_2. Sin embargo, no se utilizaría la propiedad de que los vectores son ortogonales en todo el espacio, con lo cual la búsqueda de γ_1 ($y\ \gamma_2$) es independiente de β. Estos 3 valores son codificados utilizando una tabla de ganancia y un valor que representa la "energía del frame" [12].

El decodificador toma los datos enviados desde el codificador y, al igual que todos los demás de la familia CELP, rehace la secuencia con las siguientes excepciones:

- Los coeficientes para la síntesis LPC son los "originales" (no los expandidos).
- No hay ciclo de búsqueda en base al error (close-loop).
- Hay un filtro adaptivo para la señal de salida.

3.3 Codificación basada en Redes Neuronales Artificiales

El sistema se compone esencialmente de cuatro partes:

- Compresor: encargado de identificar toda la información redundante en la señal para luego generar los parámetros que se van a transmitir.
- Transmisor: contiene un codificador diseñado para hacer un uso eficiente de las características estacionales de los parámetros generados en la etapa anterior.
- Receptor: recibe los parámetros codificados desde el canal de transmisión e incorpora un

decodificador.

- Descompresor: sintetiza la señal de voz, a partir de los parámetros recibidos.

3.3.1 Los parámetros esenciales

La idea de la compresión de voz supone que existen formas básicas cuya combinación es capaz de crear un trozo de señal más complejo. Estas formas básicas se caracterizan a través de lo que se denominó parámetros esenciales.

Tanto en el transmisor como en el compresor existen sendos arreglos (codebooks) que contienen las formas básicas (codewords) del habla. La compresión sólo consiste en encontrar la posición de los codewords que definen la señal y transmitir su ubicación en los arreglos.

Para lograr esto fueron desarrollados [1][11] cinco módulos encargados de transmitir los parámetros esenciales. Estos son (Figura 6):

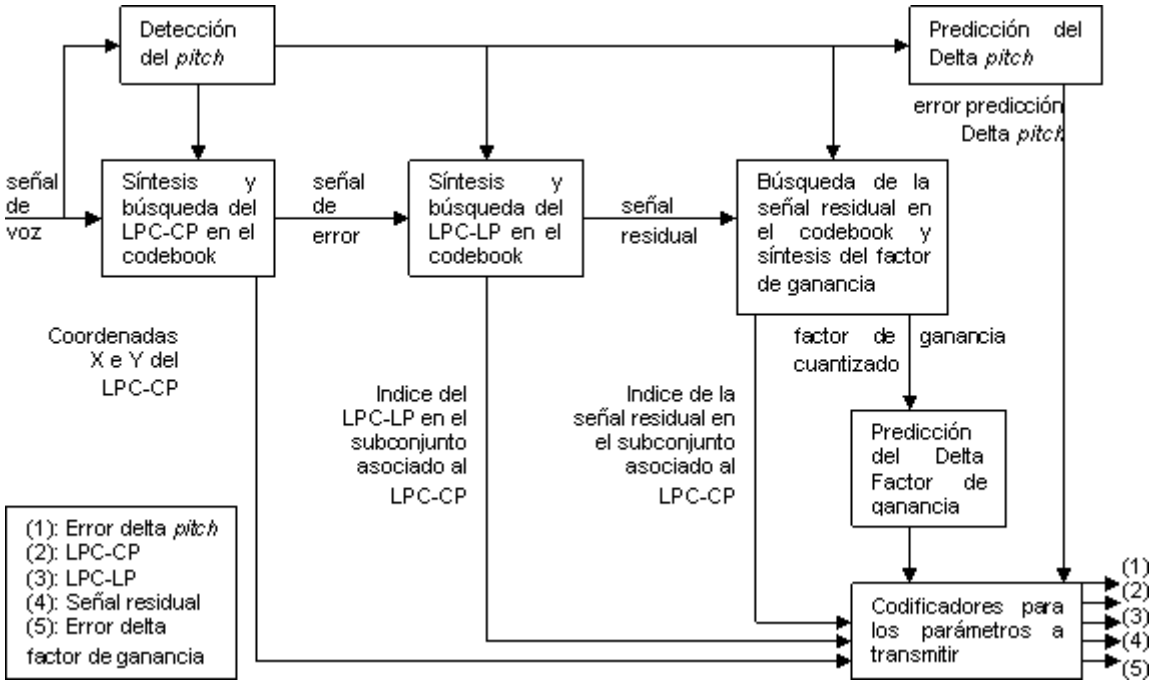


Figura 6: Ensamblaje de los módulos del codificador

- Detector de pitch. Predición de éste utilizando redes neuronales retropropagativas.
- Síntesis de LPC de corto plazo y búsqueda en el codebook. El codebook está formado por los centroides de una red neuronal autoorganizativa de Kohonnen.
- Síntesis de LPC de largo plazo y búsqueda en el codebook.
- Síntesis de señal residual y búsqueda en el codebook.
- Transmisión del factor de ganancia.

Se utiliza una cuantización de 32 niveles, normalmente codificadas en 5 bits, pero aprovechando que el factor ganancia es un parámetro que cuantizado tiene variaciones suaves en el tiempo se utiliza la transmisión de la diferencia de ganancia, codificada en 3 bits.

3.4 GSM 06.10 RPE-LTP

(Regular Excitation Long-Term Predictor)

Este es el sistema de compresión utilizado por la telefonía celular europea, utiliza la tecnología LPC para realizar la predicción de largo plazo.

La entrada de GSM 06.10 consiste en ventanas de 160 valores PCM [KHz], codificadas a 13 bits con signo. Cada ventana es de 20 ms, lo cual es alrededor de un período glotal para una voz grave o 10 para voces agudas. Este tiempo es bastante corto y durante él la señal de voz no cambia demasiado.

El compresor GSM 06.10 (Figura 7) modela el habla con dos filtros y una excitación inicial. Un filtro de predicción lineal de corto plazo; el primer paso en la compresión (y la última en la descompresión) y asume el rol del tracto vocálico y nasal. Éste es excitado por la salida de un segundo filtro de predicción lineal de largo plazo que transforma su entrada (la señal residual) en una mezcla de onda glotal y ruido.

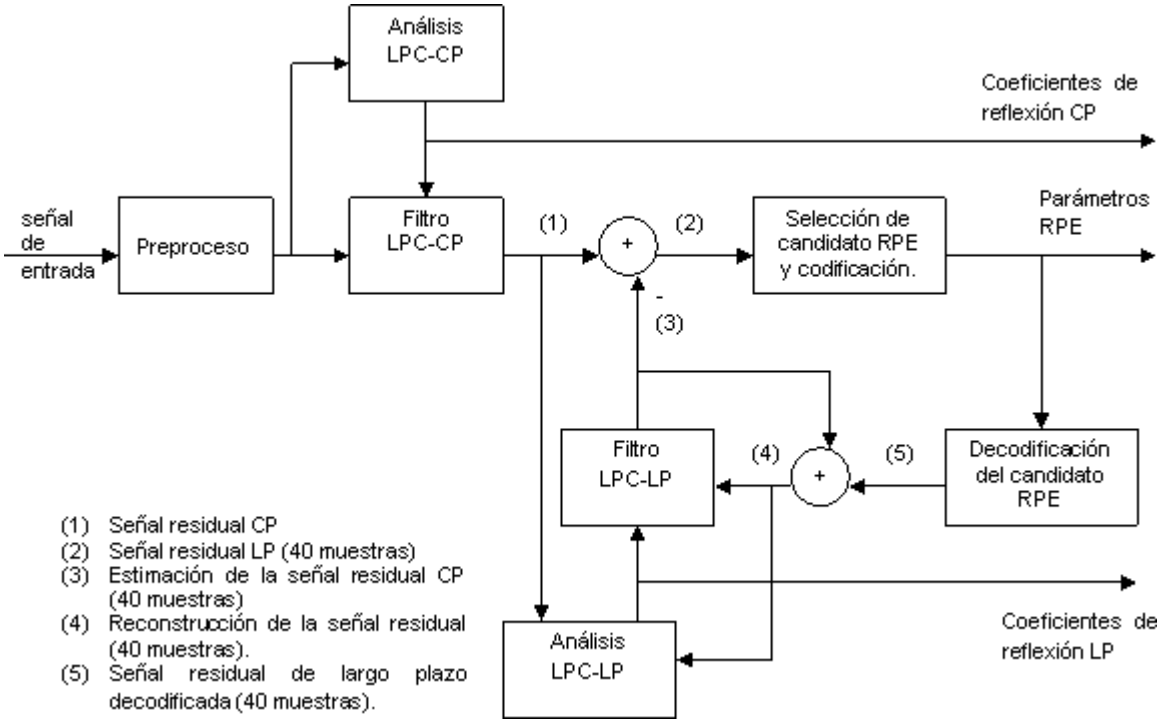


Figura 7: Esquema del codificador utilizado en GSM.

El bloque de 40 muestras de la señal residual de largo plazo es representada como una de 4 subsecuencias candidatas de 13 pulsos. La subsecuencia elegida es identificada por su posición (M) en la matriz RPE (matriz que contiene a los 4 candidatos).

El algoritmo elige la secuencia de mayor energía, esta es, la de mayor suma cuadrática de sus valores. Un índice de 2 bits transmite la elección al decodificador. Eso deja 13 muestras de 3 bits y un factor de escala de 6 bits que transforma la codificación PCM en APCM (Adaptive PCM, el algoritmo se adapta a la amplitud total aumentando o disminuyendo el factor de escala).

Finalmente, el codificador prepara la siguiente predicción a largo plazo actualizando su "salida pasada", es decir, la señal residual de corto plazo reconstruída. Para asegurarse de que el codificador y el decodificador trabajan sobre la misma señal residual, el codificador simula los pasos de progresión del decodificador hasta momentos antes de la etapa a corto plazo.

4 Estudio comparativo de la calidad de los algoritmos de compresión

El siguiente es un estudio comparativo a pequeña escala con el fin de mostrar que los sistemas de compresión de voz descritos en el capítulo 3 se encuentran en pleno funcionamiento y, además, para servir de base metodológica a posteriores estudios relacionados.

Para realizar este estudio fueron definidas dos métricas de comparación de los algoritmos, éstas son:

- Capacidad de medio de transmisión necesaria para cada algoritmo.
- Calidad de reconstrucción de la señal.

La idea de este estudio es ordenar los algoritmos según las métricas descritas sin agregar juicios de valor, éstos y la ponderación de cada métrica le corresponderá hacer a quien utilice este estudio.

4.1 Capacidad de medio de transmisión

Este punto se refiere a la capacidad mínima que debe poseer el medio de transmisión para que toda la información que envía el codificador llegue sin problema alguno (pérdida, traslape, error) al decodificador y para que ésta transmisión sea en tiempo real.

Su unidad de medida es en Kilobits por segundo, un estándar para transmisión de redes y telecomunicaciones e indica la cantidad de miles de bits que se transmiten en esa unidad de tiempo, en desmedro de cuantos bits sean necesarios para representar un símbolo (señal).

La siguiente tabla proporciona la capacidad de los distintos algoritmos:

Algoritmo	RNA	CELP	VSELP	GSM
[Kbps]	2.4	4.3	8.0	13.0

Estos valores son los teóricos, RNA fue obtenido de [11], CELP de [10], VSELP de [12] y GSM de [13].

4.2 Test MOS

El Test MOS consiste en una evaluación subjetiva de la calidad de síntesis de voz de un sistema. Fue normalizado por el comité Consultivo Internacional de Telefonía y Telegrafía (CCITT) a principio de los años 80 y se le ha utilizado principalmente para medir la calidad en sistemas de comunicación celular digital.

El test consiste en realizar una encuesta de opinión a un conjunto de individuos de prueba los cuales deben evaluar una grabación de voz según la siguiente tabla:

NOTA	CALIDAD	ESFUERZO DE ESCUCHA	DEGRADACIÓN
5	Excelente	Posible relajación completa, no requiere ningún esfuerzo.	Inaudible
4	Buena	Atención necesaria, no se requiere esfuerzo apreciable.	Audible pero no molesta.
3	Aceptable	Se necesita esfuerzo moderado.	Ligeramente molesta.
2	Mediocre	Se necesita esfuerzo considerable	Molesta.
1	Mala	Cualquier esfuerzo no permite comprender	Muy Molesta.

Para el estudio se desarrolló una página WEB con ejemplos de frases en español que además se codificaron y luego reconstruyeron con los sistemas CELP, VSELP y GSM. No se realizó para el codificador basado en RNA debido a que ya un estudio sobre ello [11] aporta el dato buscado.

Para el desarrollo de este experimento se tomaron 14 frases (7 de hablante femenino y 7 masculino). Éstas se encuentran en su forma original y la reconstrucción de ellas realizada por los 3 algoritmos ya mencionados.

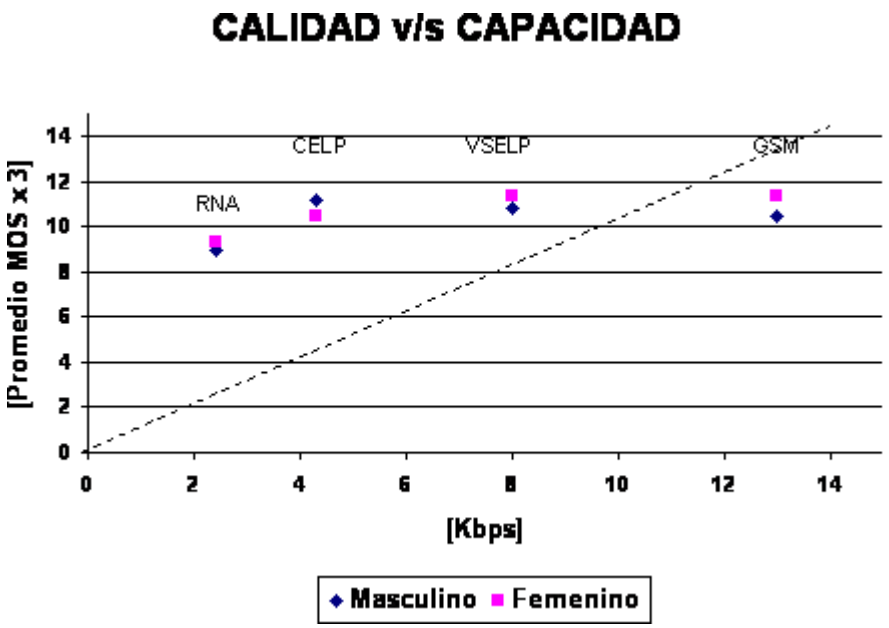
El test fue realizado por 34 personas no individualizadas, los resultados son el promedio de las notas obtenidas por frases para un hablante masculino y otro femenino; y son los siguientes:

	RNA	**CELP**	**VSELP**	**GSM**
Masculino	3.0	3.4	3.6	3.5
Femenino	3.1	3.5	3.8	3.8

Se debe notar que el valor del test MOS para RNA es en realidad su valor obtenido de [11] ponderado por un factor de corrección de apreciación que experimentalmente se obtuvo y cuyo valor es 1.28. Este factor de corrección es necesario puesto que el test aplicado en [11] se efectuó en condiciones de escaso ruido ambiente y total concentración en la percepción de la reconstrucción. Para lograr este factor, se repitió el test MOS por internet para una muestra más pequeña de ejecutantes (6 individuos) y bajo esas condiciones, se dividió el promedio resultante con los obtenidos anteriormente y se promedió esos valores para obtener el factor.

Lo más importante en este estudio es la relación entre calidad de reconstrucción del algoritmo y capacidad necesaria para la transmisión, eso se puede apreciar en el gráfico siguiente; en el cual la línea segmentada en el gráfico indica la relación 1:1 entre la nota promedio obtenido por el algoritmo en el test MOS (Calidad de reconstrucción), la cual ha sido multiplicada por un factor de escala (3) para que visualmente se aprecie mejor el efecto; versus la capacidad del medio para transmitir la codificación (medida en Kbps).

La importancia de esta relación se basa en que es inútil un algoritmo que logre una gran compresión si la calidad de la señal se pierde, y viceversa: un algoritmo que pueda reconstruir la señal de forma idéntica a la original, pero que posea una mala compresión.



5 Conclusiones

De la realización de este estudio se llega a las siguientes conclusiones:

- Se ha desarrollado una referencia a 4 sistemas de compresión de voz, sin adentrarse en los detalles de ejecución de cada uno de ellos sino, más bien, se otorga una explicación

del procedimiento y tratamiento que realizan a la señal de voz para su codificación, transmisión y posterior decodificación. Para mayor información respecto a cada uno de ellos puede consultar la bibliografía.

- Como resultado de la experimentación se aprecia que el algoritmo de mejor rendimiento (calidad de señal v/s capacidad de medio de transmisión) para codificación de voz es el algoritmo basado en Redes Neuronales Artificiales. Esto se debe a que el algoritmo envía sólo las diferencias de la información con respecto al paso anterior y aprovecha las potencialidades de las RNA para un rápido procesamiento de esa información una reconstrucción de calidad. Sin embargo, este sistema de codificación no puede ser tomado como un estándar debido a que es orientado al hablante: esto significa que el resultado del proceso depende demasiado del entrenamiento que se realice a las distintas RNA (como se demuestra en [11]).

- Por lo tanto, el algoritmo estándar de mejor rendimiento pasa a ser CELP, lo cual ha quedado demostrado al ser éste elegido como algoritmo de codificación de la telefonía celular digital norteamericana. Esto da indicios que el estudio de ésta área no ha avanzado en más de 20 años (CELP fue ideado en la década del '80) y que durante ese tiempo las investigaciones sólo se han dedicado al mejoramiento de éste sistema en vez de buscar un nuevo modelo que logre además de una mejora sustancial en rendimiento del VOCODER.

- El gran problema que posee el algoritmo CELP es el tratamiento de la señal residual, por eso la mayoría de las investigaciones realizadas se orientó a atacar ese punto y los algoritmos que de ahí se produjo, como VSELP y RPE-LTP (GSM 06.10) son una sólo una variación de CELP. Sin embargo, ellos atacan el problema de la codificación de la señal residual y no de su modelamiento.

- Este estudio intenta mejorar la forma de ejecutar el TEST MOS que la empleada en [1] y [11], dado que con la utilización de una página WEB se llega a una mayor cantidad de gente para realizarlo (34 por internet sobre 5 en papel), con una muestra más heterogénea y en las condiciones que realmente se utilizará después: un ambiente sometido a ruido e interferencias auditivas.

6 Bibliografía

[1] Velásquez Silva, Juan Domingo; Bassi Acuña, Danilo Francisco. "*Aplicaciones avanzadas de redes neuronales al desarrollo y simulación de un sistema compresor digital de voz*". Universidad de Chile, Departamento de Ingeniería Eléctrica. 1996.

[2] Mondaca Silva, Daniel Alejandro; Bassi Acuña, Danilo Francisco. "*Desarrollo y simulación de un sistema compresor de señales de voz utilizando redes neuronales*". Universidad de Chile, Departamento de Ingeniería Eléctrica. 1996.

[3] Freeman, James A.; Skapura, David M. "*Redes Neuronales: Algoritmos, Aplicaciones y técnicas de programación*". Addison-Wesley Iberoamericana, S.A. Wilmington, Delaware. E.U.A. 1993.

[4] Robinson Tony. "*Speech Analysis"*. Apuntes de Curso. 1998. http://www.dcc.uchile.cl/~abassi/WWW/Voz/Robinson98.ps.gz. Obtenido el 24/05/2000.

[5] Cawley, G.C. "*Speech Production and Perception*". 1996. http://www.dcc.uchile.cl/~abassi/WWW/Voz/Cauley96.ps.gz. Obtenido el 12/06/2000.

[6] Scilab Group, INRIA Meta2 Project/ENPC Cergrene. "*Signal processing with Scilab*". 1998. http://www.dcc.uchile.cl/~abassi/WWW/Voz/signal.ps.gz. Obtenido el 12/06/2000.

[7] Rabiner, Lawrence; Juang, Biing-Hang. "*Fundamentals of Speech Recognition*". 1993.

[8] Cerda Gazmuri, Camilo. "*Evaluación técnico-económica de una tecnología de compresión de voz digital para telefonía móvil*". Universidad de Chile, Departamento de Ingeniería Industrial. 2000.

[9] Degener, Jutta. "*Digital Speech Compression: Putting the GSM 06.10 RPE-LTP algorithm to work*". http://www.ddj.com/articles/1994/9412/9412b/9412b.htm. Obtenido el 18/08/2000.

[10] Langi, Grieder, Kinsner. "*Fast CELP Algorithm and Implementation for Speech Compression*". http://warp.ampr-umars.umsu.umanitoba.ca/umars/research/celp1.ps. Obtenido el 16/10/2000.

[11] Velásquez Silva, Juan Domingo. "*Análisis fino del tracto vocal basado en filtros LPC aplicado al mejoramiento de la calidad de síntesis de voz*". Universidad de Chile, Departamento de Ciencias de la Computación. 1996.

[12] Macres, Jason Victor. "*Theory and Implementation of the Digital Cellular Standard Voice Coder: VSELP on the TMS320C5x. Application Report*". DSP Software Engineering, Incorporated. SPRA 136. Octubre 1994.

[13] European Telecommunication Standard. "*Digital cellular telecommunication system; Full rate speech; Transcoding (GSM 06.10 version 5.0.1)*". ETSI TC-SMG, Referencia DE/SMG-110610Q. Mayo 1997.