# Sistemas de compresión de voz para telefonía celular

Alejandro Bassi A. - Javier Bustos J.
({abassi,jbustos}@dcc.uchile.cl)

Palabras clave: CELP, VSELP, GSM, RPE-LTP, LPC, Codificación de voz

Resumen

El siguiente trabajo presenta un estudio comparativo entre cuatro sistemas de compresión de voz orientados a telefonía celular: CELP, VSELP, GSM 06.10 y uno basado en Redes Neuronales Artificiales (RNA).

El principal resultado es que el sistema basado en RNA supera en rendimiento al mejor de los estándares actuales (CELP). Sin embargo, su dependencia al hablante condiciona su estandarización.

## 1. Introducción

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)[11][6][3][1], el cual necesitaba de 6.4 Kbps de capacidad en el medio de transmisión y generaba una calidad de seal 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 contaba 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 y práctica de tres sistemas de compresión utilizados en la actualidad: CELP[4], VSELP[5] y GSM 06.10[2][9]. Posteriormente se realizó un estudio comparativo de los cuatro sistemas de compresión de voz y se entregan las conclusiones de dicho estudio.

## 2. Antecedentes

A continuación se presenta el marco teórico en el cual se basan 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 el concepto de filtro y pot último 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 $cm^2$"[8]. En la Figura 2.1 se 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 del aire.

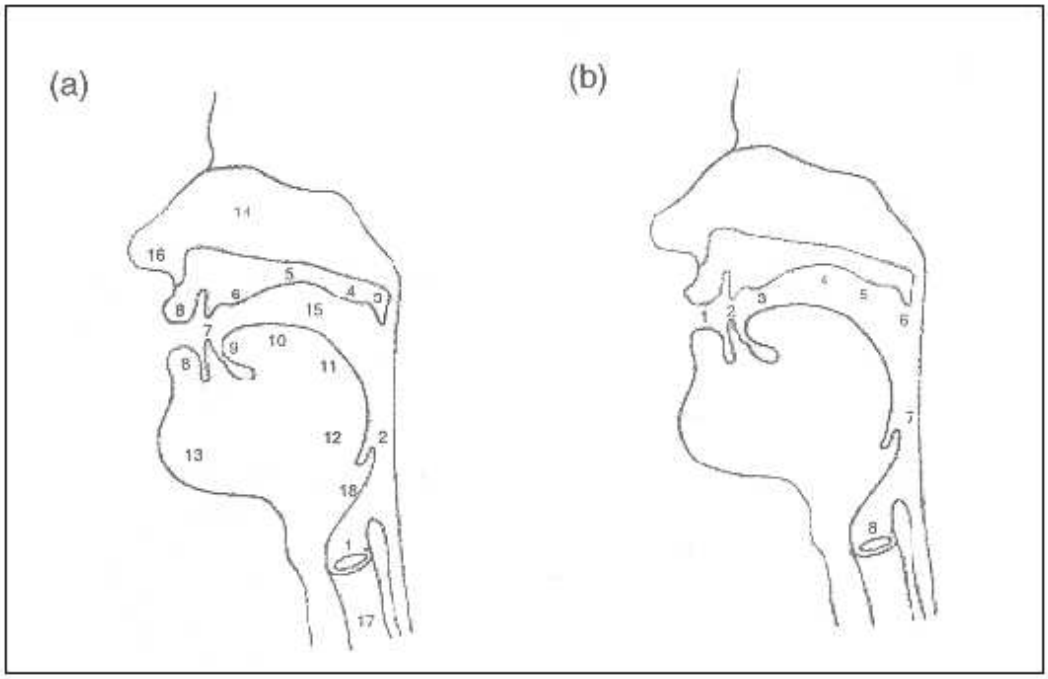


Figura 1: [7] Tracto vocálico. a) Articulaciones del habla: (1)cuerdas vocales; (2)faringe; (3)velo; (4)paladar blando; (5)paladar duro; (6)alveolos; (7)dientes; (8)labios; (9)punta de la lengua; (10)cuerpo lingual; (11)dorso; (12)raíz; (13)mandbula; (14)cavidad nasal; (15)cavidad oral; (16)ventanas 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.

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 las 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 produceel 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: generacin 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 *f*iltros 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 *r*uido blanco Figura 2.2.

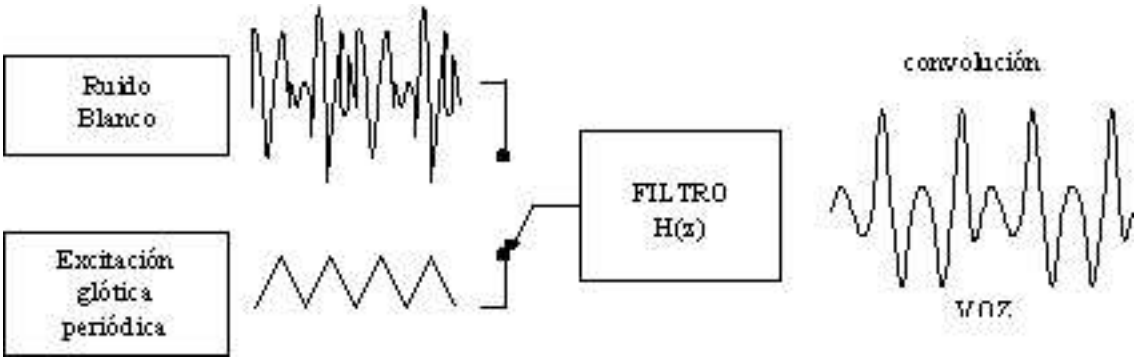


Figura 2: La voz se produce del filtrado de la excitación glótica y el ruido blanco

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 los parámetros de un filtro lineal

Para el análisis (y síntesis) de la voz se utiliza el modelo *f*uente/filtro[8][1], 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.

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:

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

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 $\hat{s}(n)$:

$$e(n) = s(n) - \hat{s}(n) \tag{2}$$

Los factores que otorgan el peso $\alpha_k$, son encontrados al minimizar el error cuadrático medio (E) calculado en una ventana de N muestras:

$$E = \frac{\sum_{i=1}^{N-1} e^2(i)}{2} \tag{3}$$

El LPC viene a representar la envolvente espectral (timbre) de la voz y el error a la excitación (prosodia).

# 3. Algoritmos de compresión de voz

En este capítulo se estudiarán cuatro algoritmos de compresión de voz 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 distorción 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 distorción es percibido de distinta manera si es aplicado a señales de voz con distinta energía y bandas de frecuencia.

La solución de CELP[4] a ese problema es utilizar la aproximación *a*nálisis por síntesis: Se realiza una síntesis de la seal, ajustando tanto los parámetros del filtro como las señales de excitación mediante un análisis comparativo con respecto a la señal original.

En la práctica, los sistemas CELP emplean algoritmos rápidos de búsqueda explotando la estructura computacional de éste. Su esquema es el siguiente(Figura 3):

- Los parámetros de la envolvente del espectro son cuantizados y representados por un set de Linear Spectrum Pairs (SLSP).
- Para el clculo de pitch se utiliza una autocorrelación de la señal original, el menor valor con el menor error será elegido y construirá una señal a partir de un codebook llamado *a*daptive codebook (ACB).
- El codebook de ruido blanco (señales de distribución Gaussianas) es llamado *s*tochastic codebook (SCB).
- Se utiliza además un filtro de percepción ($P_W$) que amplifica la señal de error en los lugares en que 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.

El decodificador toma los parámetros codificados y utilizando el mismo esquema pero en sentido inverso, reconstruye la señal $\hat{s}$. Además, se encarga de sincronizar la señal construída a partir del ACB, para ello utiliza las dos últimas muestras del subframe anterior.

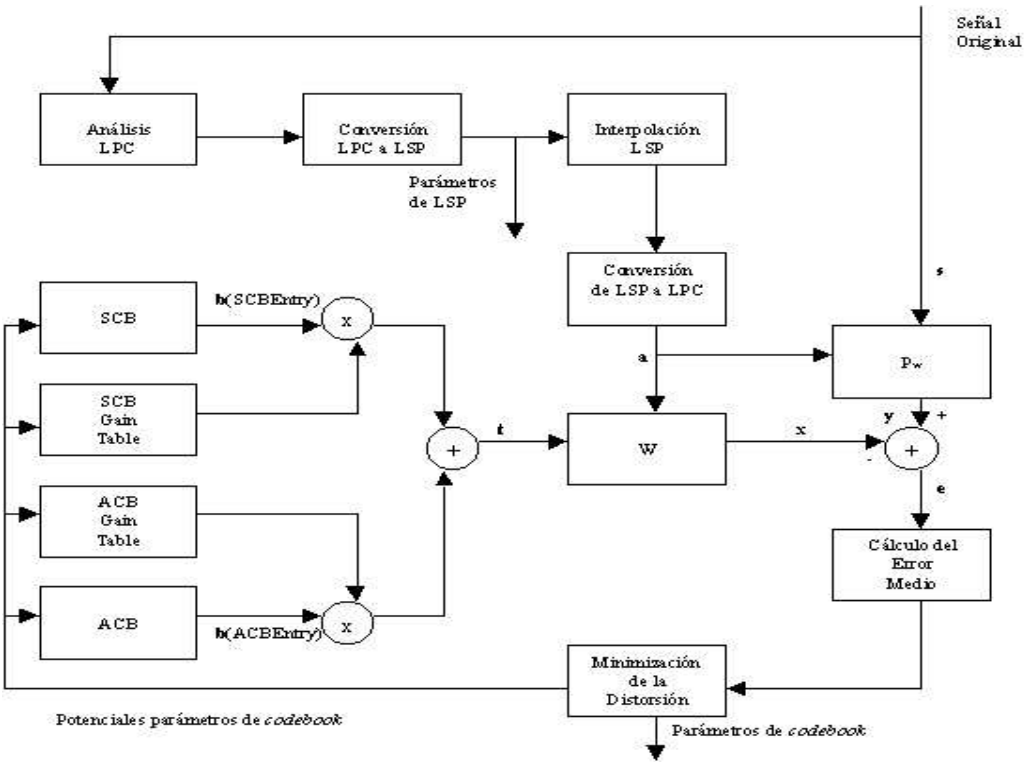


Figura 3: Esquema de un analizador CELP

## 3.2. VSELP

VSELP fue diseñado por Motorola quien es el responsable del diseño y desarrollo del algoritmo. Es una variación de CELP que codifica a 7950 bps, utilizando un adicional de 5050 bps para control de errores y sincronización de tramas.

La diferencia entre VSELP y los codificadores comunes (CELP) radica principalmente en la estructura de suc codebook. Mientras 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[5] 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 una nálisis 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 obteniendose "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 tres codebooks se cuantiza y transmite con los tres índices del codebook al receptor.

Los principales módulos de VSELP (Figura 4) 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 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.

Los valores $\beta$, $\gamma_1$ y $\gamma_2$ son codificados utilizando una tabla de ganancia y un valor que representa la "energía del frame"[5]. 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.
- Hay un filtro adaptivo para la señal de salida.

## 3.3. Codificador basado en RNA

Este sistema se diferencia de los dos anteriores por las siguientes características:

1. Posee un codebook de LPC, no envía sus coeficientes directamente.
2. Tanto el codificador como el decodificador poseen RNA de retropropagación para predecir el siguiente pitch y factor de ganancia.
3. El codificador por tanto, envía el error de la predicción de pitch y factor de ganancia al decodificador, con lo cual minimiza el tamaño del dato a transmitir.
4. La ventana de transmisión no es de tamaño fijo, sino que sincronizada al pitch.

El codebook del punto uno es construído a partir de un mapa autoorganizativo de Kohonen de topología toroidal, el cual realiza una cuantización vectorial de los coeficientes LPC. Esto produce además que la codificación LPC quede ordenada, con lo cual una pequeña variación de los parámetros produce un LPC parecido al original.

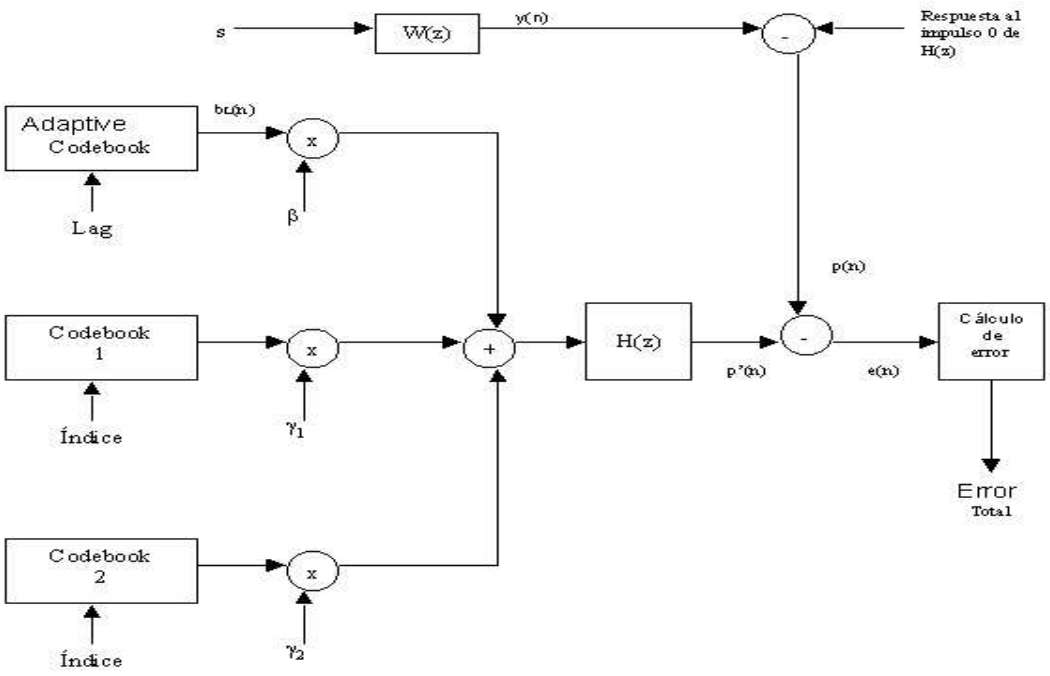


Figura 4: Esquema de un analizador VSELP

### 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 *p*arámetros esenciales.

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

Se desarrollaron cinco módulos[11][10] encargados de transmitir los parámetros esenciales (Figura 5):

- Detector de pitch. Predicción de éste utilizando redes neuronales de retropropagación.
- Síntesis de LPC de corto plazo y búsqueda en el codebook (formado por los centroides de la red autoorganizativa de Kohonen).
- Síntesis de LPC de largo plazo y búsqueda en el codebook.
- Síntesis de la señal residual y búsqueda en el codebook.
- Cuantización del factor de ganancia.

El factor de ganancia normalmente se codifica a 5 bits, pero aprovechando que éste es un parámetro que cuantizado tiene variaciones suaves en el tiempo se utiliza la transmisión de la *d*iferencia de ganancia, codificada en 3 bits.

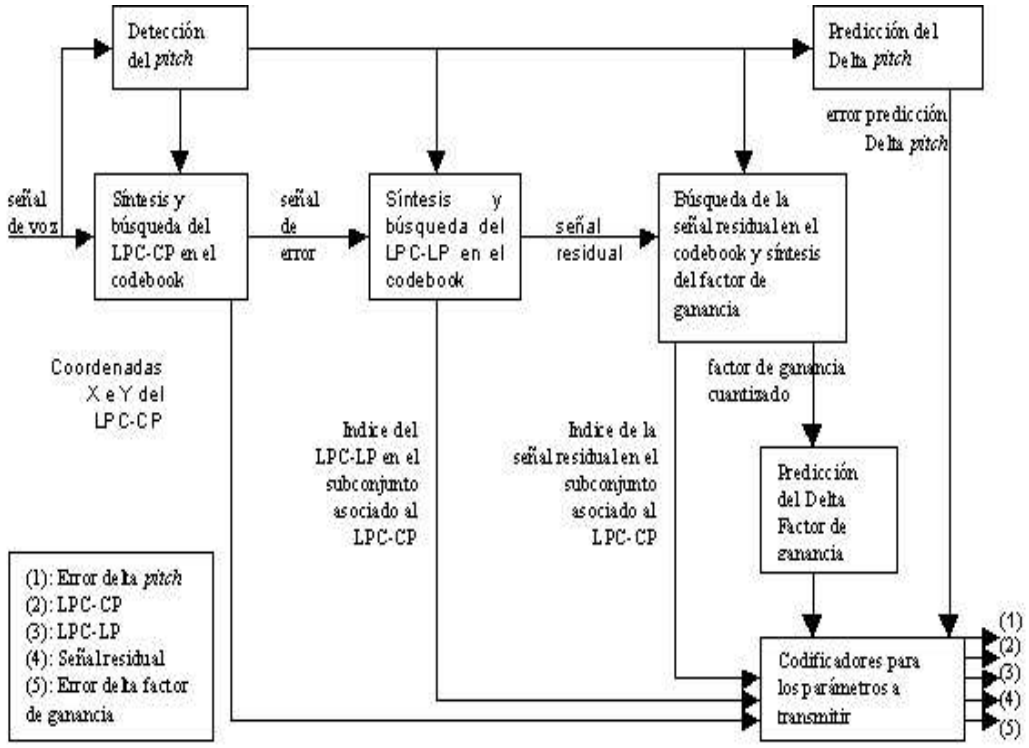


Figura 5: Ensamblaje de los módulos del codificador basado en RNA

## 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 síntesis de voz.

La entrada de GSM 06.10[2][9] consiste en ventanas de 160 valores PCM, codificadas a 13 bits con signo; cada ventana es de 20 ms. El codificador comprime la entrada de 160 valores a una ventana de 160 bits.

El compresor GSM 06.10 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) asume el rol del tracto vocálico y nasal. Este es excitado por la salida de un segundo filtro de predicción lineal de largo plaso que transforma su entrada (la señal residual) en una mezcla de onda glotal y ruido (Figura 6).

Un 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 en la matriz RPE (matriz que contiene 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 (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 que el codificador y el decodificador trabajan sobre la misma señal residual, el codificador simula los pasos de progresioón del decodificador hasta momentos antes de la etapa a corto plazo.

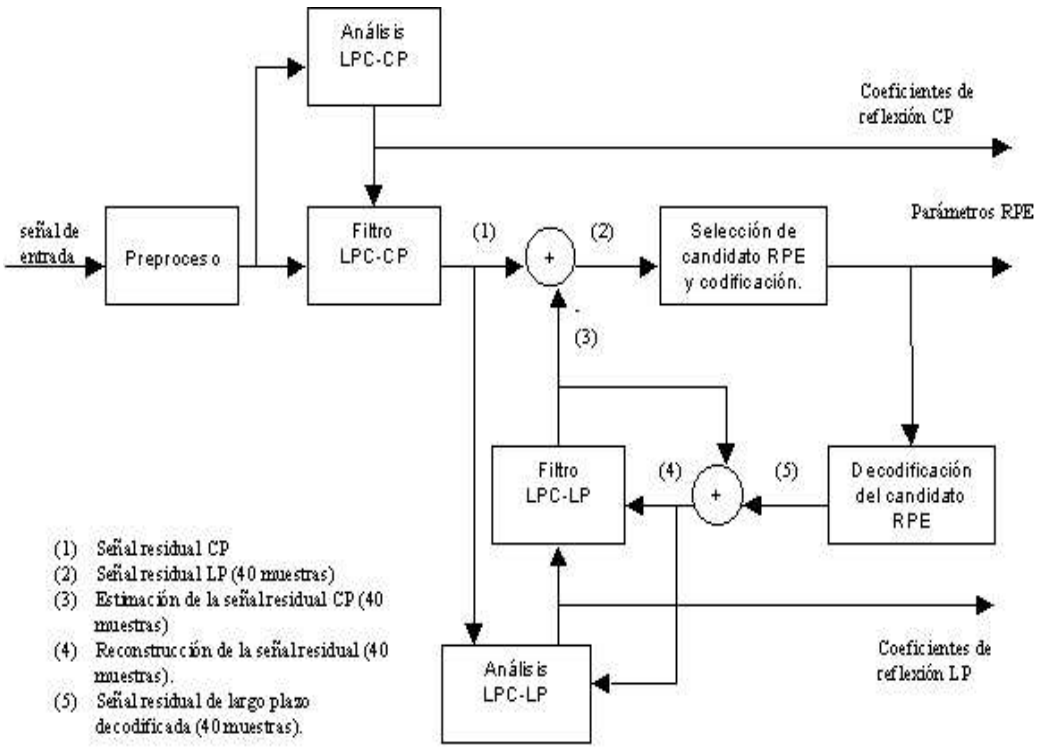


Figura 6: Esquema de un codificador utilizado en GSM 06.10

# 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 *m*étricas de comparación de los algoritmos: capacidad del medio de transmisión necesaria para cada algoritmo y calidad de reconstrucción de la señal.

## 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 (de pérdida, traslape y/o error) al decodificador y para que ésta transmisión sea en tiempo real. Su unidad de medida es en Kilobits por segundo, un estándard

para transmisior de redes y telecomunicaciones.

La siguiente tabla proporciona la capacidad de los distintos algoritmos:

| **Algoritmo** | RNA | CELP | VSELP | GSM |
|---|---|---|---|---|
| **Kbps** | 2.4 | 4.3 | 8.0 | 13.0 |

Esos valores son teóricos: RNA fue obtenido de [10], CELP de [4], VSELP de [5] y GSM de [9].

## 4.2. Test MOS

El Test MOS 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. Consiste en realizar una encuesta de opinión a un conjunto de individuos de prueba los cuales deben evaluar una grabacin segun la siguiente tabla:

| **NOTA** | **CALIDAD** | **ESFUERZO DE ESCUCHA** | **DEGRADACIÓN** |
|---|---|---|---|
| 5 | Excelente | Posible relajación completa, no requiere esfuerzo. | Inaudible. |
| 4 | Buena | Atención necesaria, mínimo esfuerzo. | 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 14 ejemplos de frases en español (7 de hablante femenino y 7 masculino) que además se codificaron y reconstruyeron con los sistemas CELP, VSELP y GSM. No se realizó para el codificador basado en RNA debido a que un estudio sobre ello[10] aporta el dato buscado.

El test fue desarrollado por 34 personas (lo que da un total de 476 notas MOS) no individualizadas, los resultados son el promedio de las notas obtenidas por frases para un hablante masculino y otro femenino; y son las siguientes:

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

Para hacer "comparable" el algoritmo basado en RNA con los otros tres su nota del test MOS fue ponderada por un factor de corrección de apreciación obtenido experimentalmente y cuyo valor es *1.*28. Para lograr este factor, se repitió el test MOS bajo las especificaciones de [10] utilizando una muestra más pequeña (se obtuvieron 84 notas MOS); luego se dividió el promedio resultante con los obtenidos anteriormente, la media de esos valores es el factor de corrección.

Lo más importante en este estudio es la relación entre la calidad de reconstrucción del algoritmo y la capacidad necesaria para la transmisión, apreciable en el gráfico de la Figura 7.

La línea segmentada en el gráfico indica la relación 1:1 entre la nota promedio obtenido por el algoritmo en el test MOS (escalada por 3 para una mejor visualización) y la capacidad del medio. 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 recosntruir la señal de forma idéntica a la original, pero que posea una mala compresión.

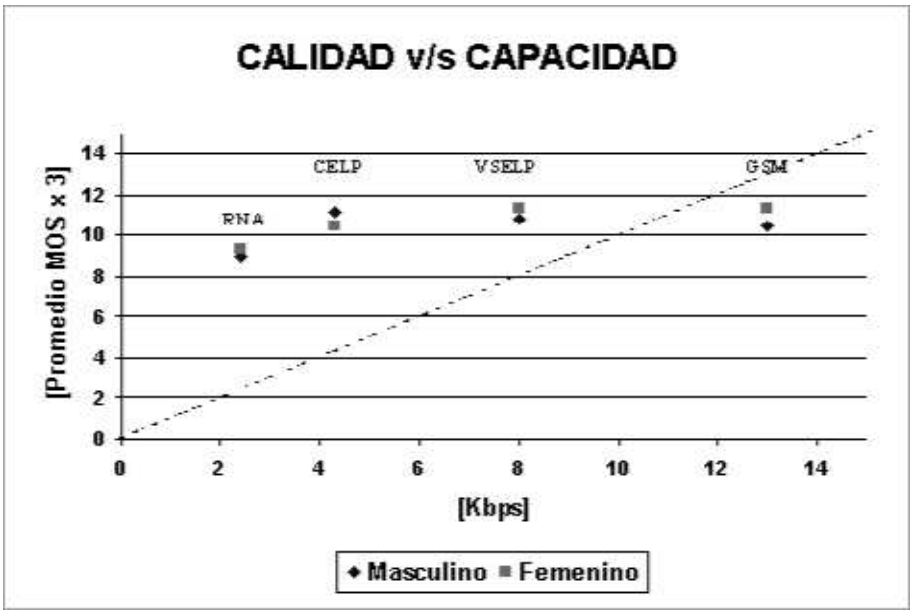


Figura 7: Gráfico de calidad v/s capacidad

# 5. 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 que se otorga una explicación del procedimiento 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 bibliografiía.

Como resultado de la experimentación se aprecia que el algoritmo de mejor rendimiento para codificación de voz es el algoritmo basado en RNA. Esto se debe a que el algoritmo envía sólo las diferencias de la información con respecto al paso anterioir y aprovecha las potencialidades de las RNA para un rápido procesamiento de esa información, logrando una reconstrucción de calidad media. Sin embargo, este sistema de codificación no puede ser estandarizado para telefona puesto a que es orientado al hablante: esto sisignifica que el resultado del proceso depende demasiado del entrenamiento que se realice a las distintas RNA (como se demuestra en [10]).

Por lo tanto, el algoritmo estándard 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 esta 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 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 nacieron (VSELP y GSM 06.10) se limitan a atacar el problema de la codificación de la señal residual y no su modelamiento. Como se puede apreciar en [8],[1] y [7], la señal residual es una señal de espectro plano orientada al pitch, y se puede modelar como pulsos (grandes amplitudes de energía en tiempos pequeños), esto lleva a pensar en la utilización de Wavelets para su modelamiento y codificación.

# Referencias


[1] G.C. Cauley. "Speech production and perception". http://www.dcc.uchile.cl/%7Eabassi/WWW/Voz/Cauley96.ps.gz, 1996.

[2] J. Degener. "Digital speech compression: Putting the gsm 06.10 rpe-ltp algorithm to work". http://www.ddj.com/articles/1994/9412/9412b/9412b.htm, 1994.

[3] J. Freeman and D. Skapura. *"Redes Neuronales: Algoritmos, aplicaciones y técnicas de programación"*. Addison-Wesley Iberoamericana S.A., Wilmington, Delawere, U.S.A, 1993.

[4] Grieder Langi and Kinsner. "Fast celp algorithm and implementation for speech compression". http://warp.ampr-umars.umsu.umanitoba.ca/umars/research/celp1.ps, 2000.

[5] J.V Macres. "Theory and implementation of the digital cellular standard voice coder: VSELP on the TMS320C5x". Application report, DSP Software Engineering, Incorporated, October 1994.

[6] D. Mondaca. "Desarrollo y simulación de un sistema compresor de señales de voz utilizando redes neuronales". Master's thesis, Departamento de Ingeniería Eléctrica. Universidad de Chile, 1996.

[7] L. Rabiner and J. Biing-Hang. *"Fundamentals of Speech Recognition"*. 1993.

[8] T. Robinson. "Speech analysis". http://www.dcc.uchile.cl/%7Eabassi/WWW/Voz/Robinson98.ps.gz, 1998.

[9] European Telecomunication Standard. "Digital cellular telecommunication system; full rate speech; transcoding (GSM 06.10 version 5.0.1)". Thecnical report, ETSI TC-SMG, May 1996.

[10] J. Velásquez. "Análisis fino del tracto vocal basado en filtros LPC aplicado al mejoramiento de la calidad de síntesis de voz". Master's thesis, Departamento de Ciencias de la Computación. Universidad de Chile, 1996.

[11] J. Velásquez. "Aplicaciones avanzadas de redes neuronales al desarrollo y simulación de un sistema compresor digital de voz". Master's thesis, Departamento de Ingeniería Eléctrica. Universidad de Chile, 1996.