CODIFICACION
DE
VOZ Y VIDEO

Dr. Ing. José Joskowicz
josej@fing.edu.uy
Instituto de Ingeniería Eléctrica, Facultad de Ingeniería
Universidad de la República
Montevideo, URUGUAY

Febrero 2017

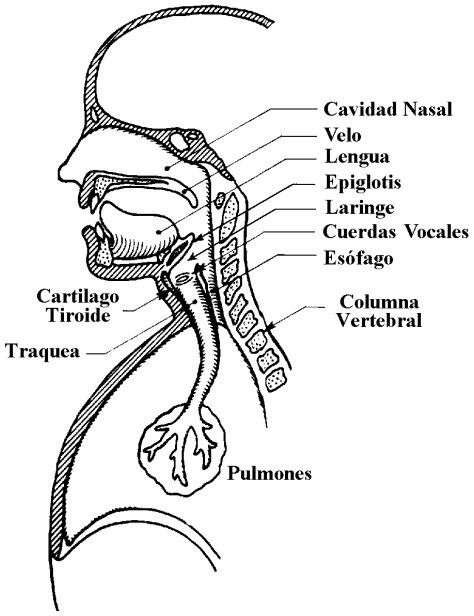
Contenido

1 Introducción a la codificación de voz

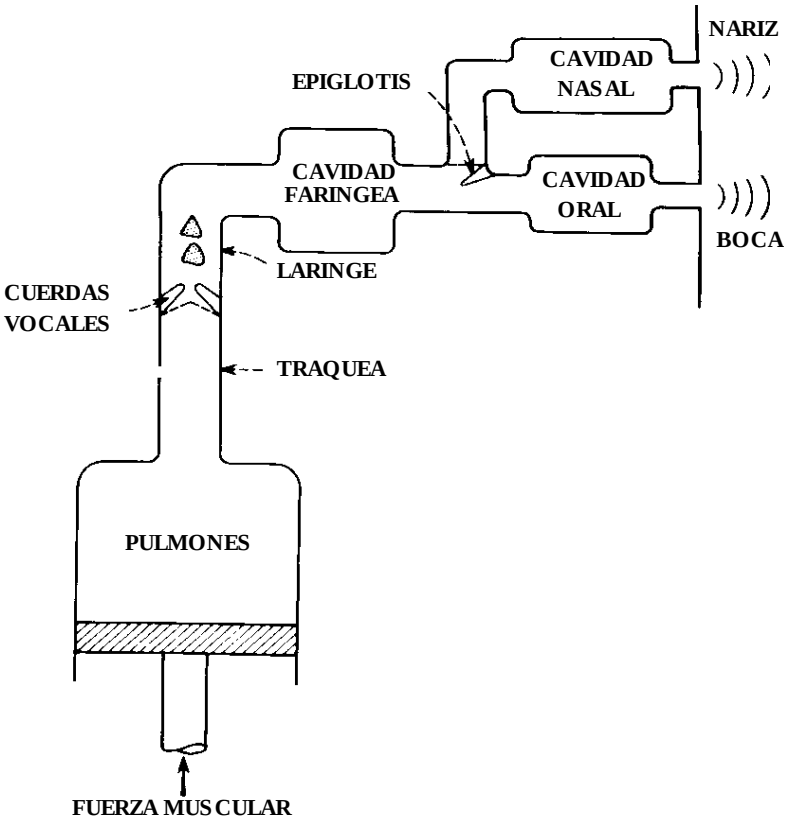
Las centrales telefónicas digitales realizan la conmutación de audio en forma digital. Las centrales IP (IP PBX, Softswitches, etc.) utilizan las redes de datos para realizar el envío de audio entre dispositivos, a través del envío de paquetes de datos. Esto requiere que en algún punto del sistema la señal de voz analógica sea digitalizada, es decir, convertida en una secuencia de número discretos. Este proceso puede realizarse en los propios teléfonos (cómo es el caso en los "teléfonos digitales" o en los "teléfonos IP"), en "Gateways" (o conversores de medios y señalización) o las "placas de abonados" entre otros.

Las primeras ideas acerca de la digitalización del audio son atribuidas al Ingeniero Alec Reeves, quien desarrolló el primer sistema de audio digital, con fines militares en 1937. La inminente segunda guerra mundial hacía necesario disponer de sistemas de transmisión telefónicas más seguros. Si bien la idea fue patentada por Reeves, su popularización debió esperar por varias décadas al desarrollo de nuevas tecnologías, (más específicamente, a la invención del transistor). La tecnología de PCM se popularizó sobre fines de la década de 1960, momento para el cual ya no eran reclamables derechos por la patente.

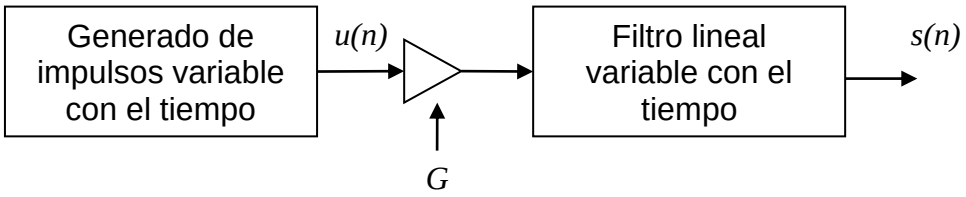
La codificación de la voz ha evolucionado notoriamente desde las primeras ideas de Alec Reeves. En [1] se puede leer una breve y esclarecedora historia de la evolución de la codificación de voz, narrada por uno de sus principales protagonistas, Bishnu S. Atal (quien diseñó las técnicas de codificación conocidas como CELP). Inicialmente, los codecs se basaron en codificar de la manera más eficiente posible la "forma de onda" de la señal, utilizando características de la voz y el oído (por ejemplo, se ha comprobado que el oído humano es más sensible a ruidos o distorsiones en señales de baja amplitud que a los mismos ruidos o distorsiones en señales de mayor amplitud). Tal es el caso de los codecs del tipo PCM (que serán descritos en detalle más adelante). Posteriormente, a los efectos de poder bajar la tasa de bits necesaria para la transmisión, se comenzaron a utilizar técnicas "predictivas". Éstas técnicas están basadas en predecir los valores de las muestras en base a la extrapolación de las muestras anteriores, y codificar únicamente la diferencia entre la predicción y el valor real de la muestra. Esta predicción puede realizarse en forma fija o adaptiva, la que logra mucho mejores resultados. Las técnicas predictivas dieron origen a la tecnología conocida como "LPC" (Linear Prediction Coding), la que fue desarrollada sobre fines de los años 1960. En 1973 fue desarrollado el primer sistema práctico que utilizó técnicas del tipo LPC.
Sobre la década de 1980, una nueva idea es introducida en la codificación de la voz. Esta idea consiste en generar "voz sintética", simulando la manera en que se produce la voz humana en el conducto vocal. La siguiente figura muestra la anatomía del aparato fonador, desde los pulmones hasta los labios.



Es posible realizar un modelo del conducto vocal basado en un generador de impulsos de aire y un conjunto de tubos, de diámetros variables, como se muestra en la siguiente figura. Los tubos en su conjunto se comportan como un filtro que varía lentamente con el tiempo, pero con propiedades estables (cuasi-estacionarias) en intervalos cortos (de alguna decena de milisegundos).



Un generador de impulsos puede modelar la señal que excita el conducto vocal. En el caso de ser una señal “sonora” (por ejemplo las vocales), el modelo es un generador de impulsos periódicos, simulando la apertura y cierre periódicos de las cuerdas vocales, a una frecuencia dada por el tono (o “pitch”) de la voz. De tratarse de un sonido sordo (por ejemplo la “m” o la “s”), el modelo es más parecido a un generador de "ruido blanco", simulando la señal luego de pasar por un estrechamiento del conducto vocal (la turbulencia del aire luego de pasar por dicho estrechamiento tiene un comportamiento muy aperiódico). El conducto vocal en su conjunto puede ser modelado como un filtro de respuesta variable en el tiempo (pero estacionario en periodos cortos), cuya excitación proviene de un generado de impulsos. Este modelo se esquematiza en la siguiente figura, donde $u(n)$ representa un tren de impulsos, G es una constante relacionada con la ganancia (o el volumen de voz del locutor) y $s(n)$ es la salida sintetizada de la voz.



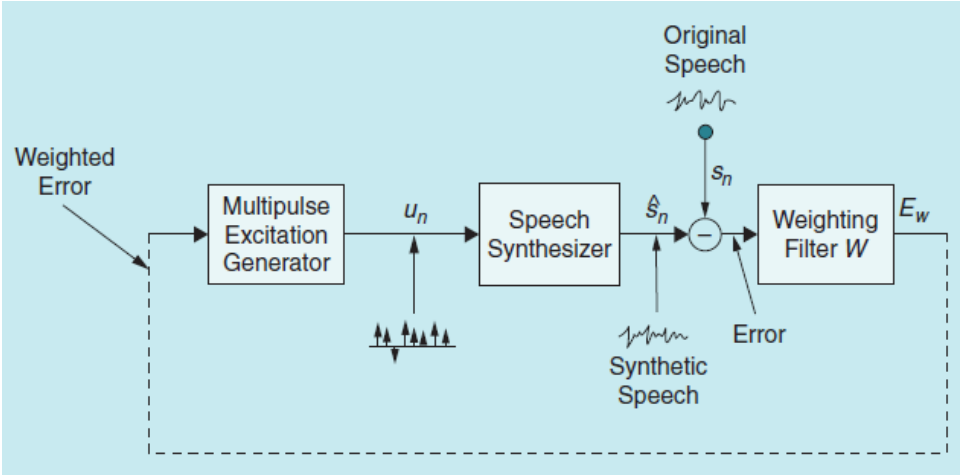
La forma matemática de la transferencia del filtro lineal, dentro de cada intervalo corto de tiempo, surge del modelo de tubos de la figura anterior, y puede ser expresada como:

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

Donde p es el orden del filtro, y a_k representan los coeficientes del filtro.

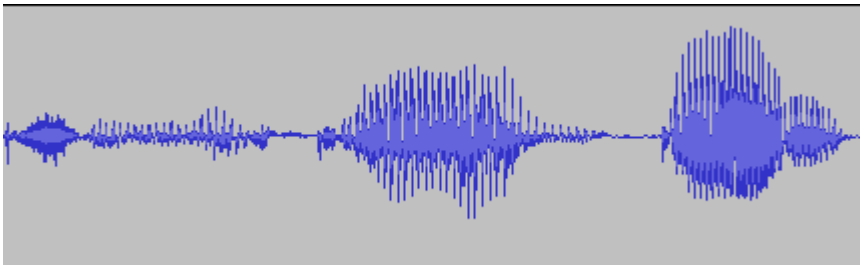
El generador de impulsos $u(n)$ está muy relacionado al “tono” o “pitch” de la voz para señales sonoras, y consiste en una serie de impulsos separados por un tiempo dado por el tono o “pitch” de la voz. Para señales sordas, el patrón de pulsos puede ser más complejo, asemejándose más a un “ruido blanco”.

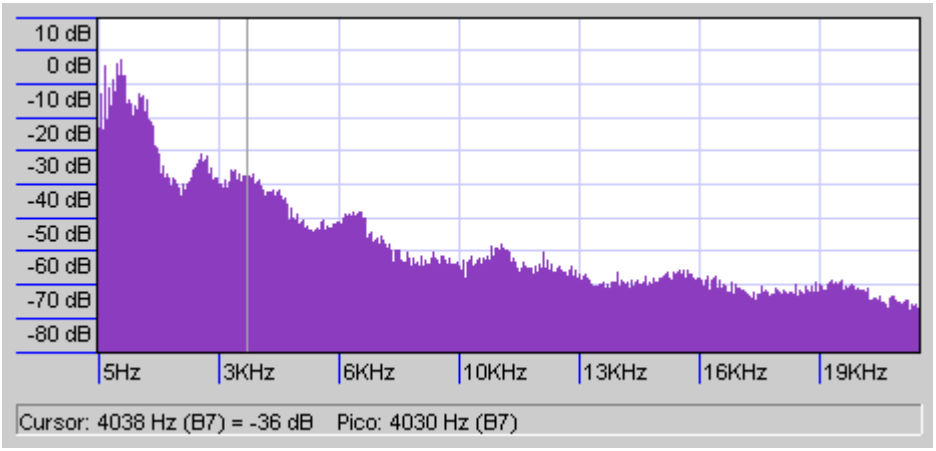
Con esto en mente, el modelo de síntesis de voz consiste en encontrar, para cada período corto de tiempo, los mejores valores de los coeficientes a_k, de la ganancia G y del generador de impulsos $u(n)$. Estos valores deben ser tales que minimicen las diferencias entre una onda de sonido real y la sintetizada, como se muestra en la siguiente figura (extraída de [1]).



Varios codecs actuales están basados en estas ideas, como se verá en las siguientes secciones.

La voz humana puede tener tonos que lleguen hasta los 20 kHz, lo que se corresponde con el límite de frecuencias que puede escuchar el oído. Sin embargo, la mayor parte de la energía de la voz se centra en las frecuencias bajas, típicamente por debajo de los 4 kHz. En la siguiente figura se muestra un segmento de audio de voz, y su corresponde espectro. Se puede ver como, luego de los 4 kHz, la energía de la señal decrece rápidamente.





2 Digitalización y codificación de la voz

2.1 CODECs

Los codecs son los dispositvos que realizan la codificación y decodificación de la voz. Pueden ser caracterizados por diferentes aspectos, entre las que se encuentran su tasa de bits (bit rates), la calidad resultante del audio codificado, su complejidad, el tipo de tecnología utilizada y el retardo que introducen, entre otros. Originalmente, los primeros codecs fueron diseñados para reproducir la voz en la banda de mayor energía, entre 300 Hz a 3.4 kHz. Actualmente este tipo de codecs son caracterizados como de "banda angosta" (narrowband). En contraste, los codecs que reproducen señales entre 50 Hz y 7 kHz se han llamado de "banda ancha" (wideband). Más recientemente, ITU-T ha estandarizado codecs llamados de banda superancha (superwideband), para el rango de 50 Hz a 14 kHz y de banda completa (fullband), para el rango de 50 Hz a 20 kHz [2]

La siguiente tabla muestra algunos de los Codecs más conocidos. Varios de ellos son detallados en las siguientes secciones. Las recomendaciones de los codecs estandarizados por ITU-T están disponibles en la página de ITU-T [3].

Codecs de banda angosta (narrowband):

Codec	Nombre	Bit rate (kb/s)	Retardo (ms)	Comentarios
G.711	PCM: Pulse Code Modulation	64, 56	0.125	Codec "base", utiliza dos posibles leyes de compresión: µ-law y A-law [4]
G.723.1	Hybrid MPC-MLQ and ACELP	6.3, 5.3	37.5	Desarrollado originalmente para video conferencias en la PSTN, es actualmente utilizado en sistemas de VoIP [5]
G.728	LD-CELP: Low-Delay code excited linear prediction	40, 16, 12.8, 9.6	1.25	Creado para aplicaciones DCME (Digital Circuit Multiplex Encoding) [6]
G.729	CS-ACELP: Conjugate Structure Algebraic Codebook Excited Linear Prediction	11.8, 8, 6.4	15	Ampliamente utilizado en aplicaciones de VoIP, a 8 kb/s [7]
AMR	Adaptive Multi Rate	12..2 a 4.75	20	Utilizado en redes celulares GSM [8]

Codecs de banda ancha (wideband):

Codec	Nombre	Bit rate (kb/s)	Retardo (ms)	Comentarios
G.722	Sub-band ADPCM	48,56,64	3	Inicialmente diseñado para audio y videconferencias, actualmente utilizado para servicios de telefonía de banda ancha en VoIP [9]

Codec	Nombre	Bit rate (kb/s)	Retardo (ms)	Comentarios
G.722.1	Transform Coder	24,32	40	Usado en audio y videoconferencias [10]
G.722.2	AMR-WB	6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85	25.9375	Estandar en común con 3GPP (3GPP TS 26.171). Los bit rates más altos tienen gran inmunidad a los ruidos de fondo en ambientes adversos (por ejemplo celulares) [11]
G.711.1	Wideband G.711	64, 80, 96	11.875	Amplía el ancho de banda del codec G.711, optimizando su uso para VoIP [12]
G.729.1	Wideband G.729	8 a 32 kb/s	<49 ms	Amplía el ancho de banda del codec G.729, y es “compatible hacia atrás” con este codec. Optimizado su uso para VoIP con audio de alta calidad [13]
RtAudio	Real Time Audio	8.8, 18	40	Codec propietario de Microsoft, utilizado en aplicaciones de comunicaciones unificadas (OCS) [14]

Codecs de banda super ancha (superwideband):

Codec	Nombre	Bit rate (kb/s)	Retardo (ms)	Comentarios
SILK	SILK	8 a 24	25	Utilizado por Skype [15]

Codecs de banda completa (fullband):

Codec	Nombre	Bit rate (kb/s)	Retardo (ms)	Comentarios
G.719	Low-complexity, full-band	32 a 128	40	Es el primer codec “fullband” estandarizado por ITU [16]

2.1.1 G.711

El codec básico y mas antiguo en telefonía es el estandarizado en la recomendación G.711 de la ITU-T [4], implementando la “ley A” o “ley μ”. Mediante esta codificación se obtiene una señal digital de 64 kb/s, como se verá a continuación.

El codec G.711 es del tipo de “forma de onda”. Cada muestra de audio es digitalizada, cuantificada y codificada, según el proceso que se describe a continuación.

2.1.2 Muestreo

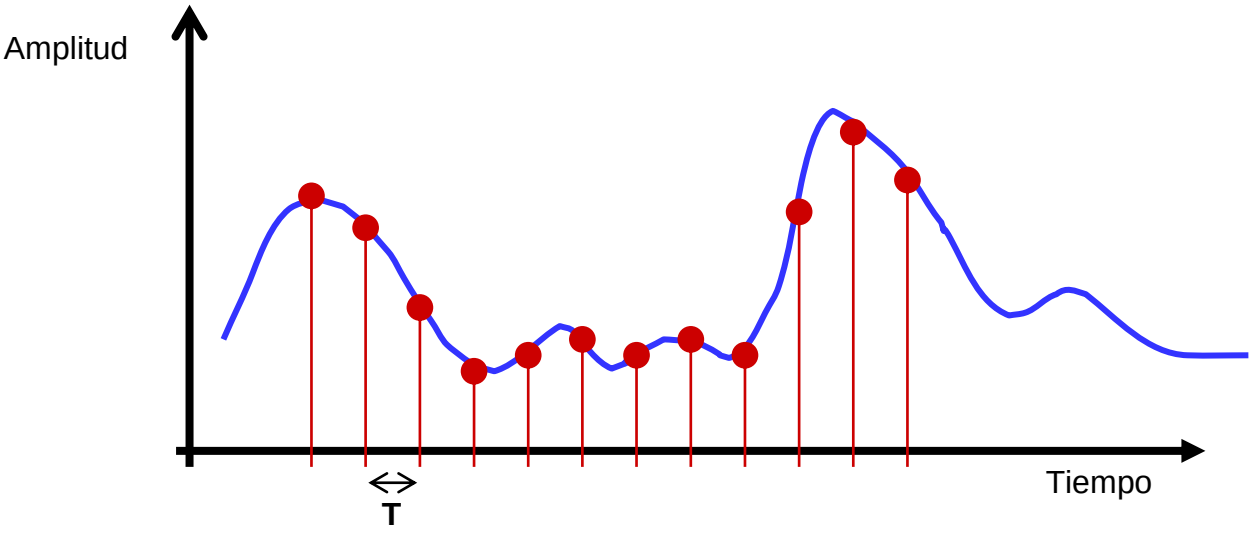
El proceso de "muestreo" consiste en tomar muestras de la señal vocal a intervalos regulares. Estos intervalos deben ser tales que cumplan con el "Teorema del muestreo", que establece:

> *"La mínima frecuencia a la que puede ser muestreada una señal y luego reconstruida sin perder información, es el doble de la frecuencia máxima de dicha señal"*

Para establecer cual es ésta frecuencia mínima en el codec G.711 se han tenido en cuenta las siguientes consideraciones de las señales de voz:

- Si bien el oído humano puede llegar a escuchar sonidos de hasta 18 a 20 kHz, la mayor parte de la energía de las señales de voz humana se encuentran por debajo de los 4 kHz.
- El sonido resultante de filtrar la voz humana a 3.4 kHz es perfectamente inteligible, y además se puede distinguir sin problemas al locutor.
- El sistema de telefonía originalmente se ha diseñado para transmitir satisfactoriamente "voz humana", minimizando los recursos necesarios para ésta tarea.

Por lo expuesto, podemos pensar en un ancho de banda mínimo para las señales de los sistemas de telefonía de 3.4 kHz. Según el teorema del muestro, para poder reconstruir una señal de hasta 3.4 kHz, debe ser muestreada a más de 6.8 kHz. Dado que los "filtros reales" no pueden realizar cortes abruptos, se ha tomado originalmente en el codec G.711 una frecuencia de muestreo de 8 kHz, es decir, tomar una muestra de voz cada 125 microsegundos. Si bien esto es adecuado para reproducir la voz humana, el audio de "alta calidad", por ejemplo con contenido de música, requiere de frecuencias de muestro mucho mayores, para que puedan llegar a funcionar con señales de hasta 20 kHz.

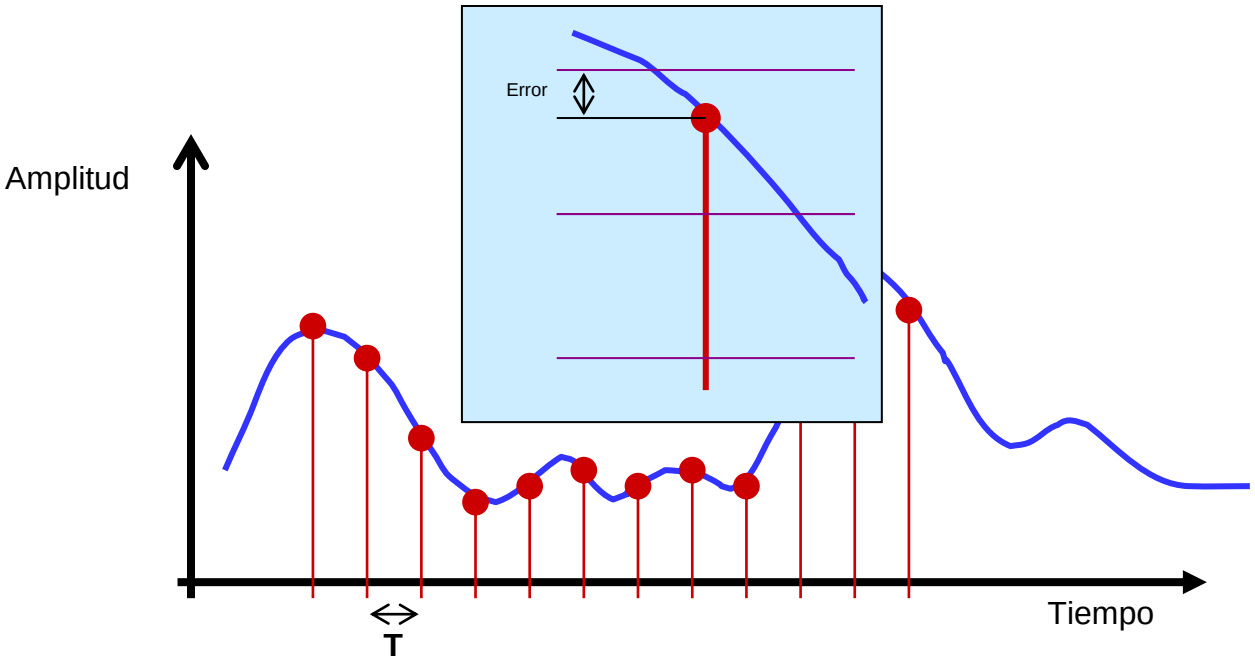


2.1.3 Cuantificación

El proceso de cuantificación convierte las muestras analógicas en muestras que pueden tomar un conjunto discreto de valores. De esta manera, los valores de las muestras se “cuantifican” en cantidades discretas.

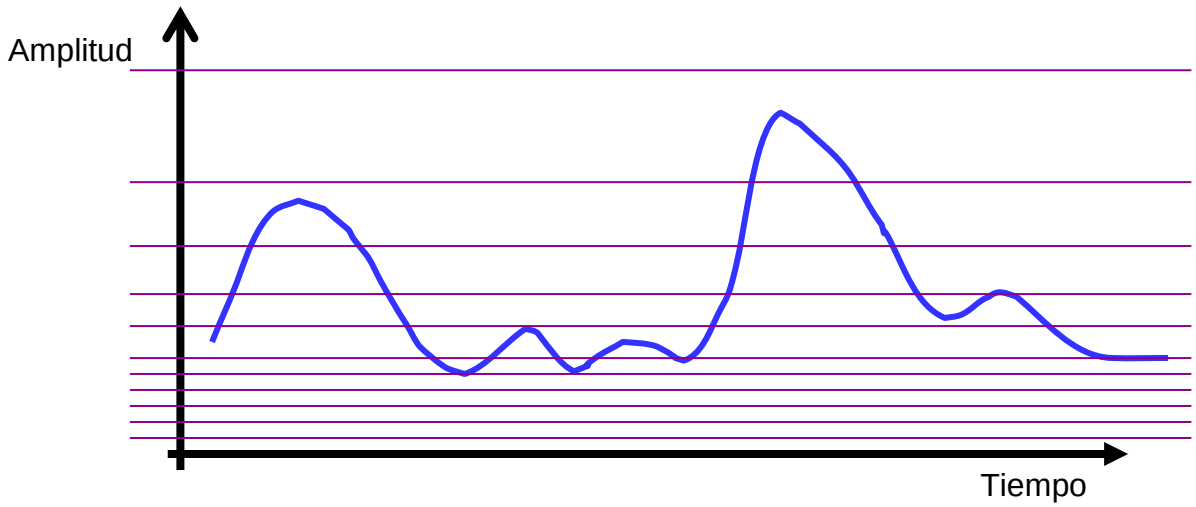
Al pasar de infinitos valores (señal analógica) a un conjunto discreto de valores, se introduce naturalmente una distorsión a la señal original. Esta distorsión se conoce normalmente como “Ruido de Cuantificación”. Es de hacer notar, que más allá de su nombre, esta distorsión no es un “ruido”, ya que no proviene de factores externos, sino que es parte del propio proceso de digitalización.

Cuántos más valores discretos se utilicen, menor será la distorsión introducida en el proceso. Por otro lado, cuántos más valores discretos se utilicen, mayor será la cantidad de “información” (bits) que se deben procesar (o transmitir) por cada muestra.



Es importante detenernos a pensar cuál es la menor cantidad de "valores discretos" aceptables para el tipo de señal que se desea digitalizar. Se ha demostrado que para lograr niveles de "ruido" aceptables al reconstruir señales de voz cuantificadas, se requieren de unos 4.000 niveles de cuantificación, utilizando una "cuantificación lineal" (esto es, dividiendo en intervalos de la misma amplitud el "eje y"). Esto requiere de 12 bits por muestra (recordar que con 12 bits se pueden representar 2^{12} valores = 4096 valores).
Por otro lado, se ha comprobado que el oído humano es más sensible a ruidos o distorsiones en señales de baja amplitud que a los mismos ruidos o distorsiones (en valores absolutos) en señales de mayor amplitud. Esto lleva a pensar en algún tipo de cuantificación no lineal, de manera de disponer de distorsiones pequeñas en las partes de baja amplitud, a costo de distorsiones mayores en las partes de gran amplitud de la señal.

El proceso de cuantificación adoptado originalmente en telefonía por la CCITT (y actualmente estandarizado por ITU en la Recomendación G.711) implementa un algoritmo no lineal, de manera de obtener una calidad de voz aceptable, minimizando la cantidad de "niveles de cuantificación". Este algoritmo se basa en tener distorsiones pequeñas para las amplitudes pequeñas de la señal, y distorsiones mayores para las amplitudes mayores de la señal.

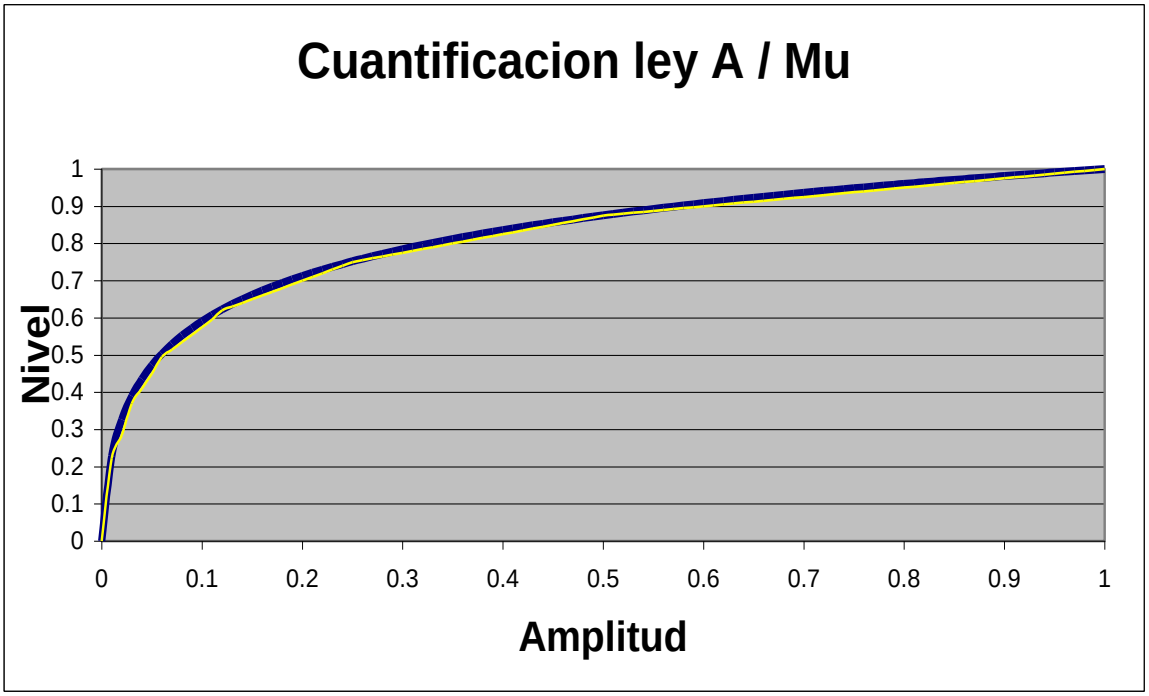


Ley A (de 13 segmentos)

$y=(1+\log(Ax) / (1 + \log(A))$ si $1/A < x < 1$
$y = Ax/(1 + \log(A))$ si $0 < x < 1/A$
$A = 87.6$

Ley μ (de 15 segmentos)

$y = \log(1+\mu x)/\log(1+\mu)$
$\mu = 255$



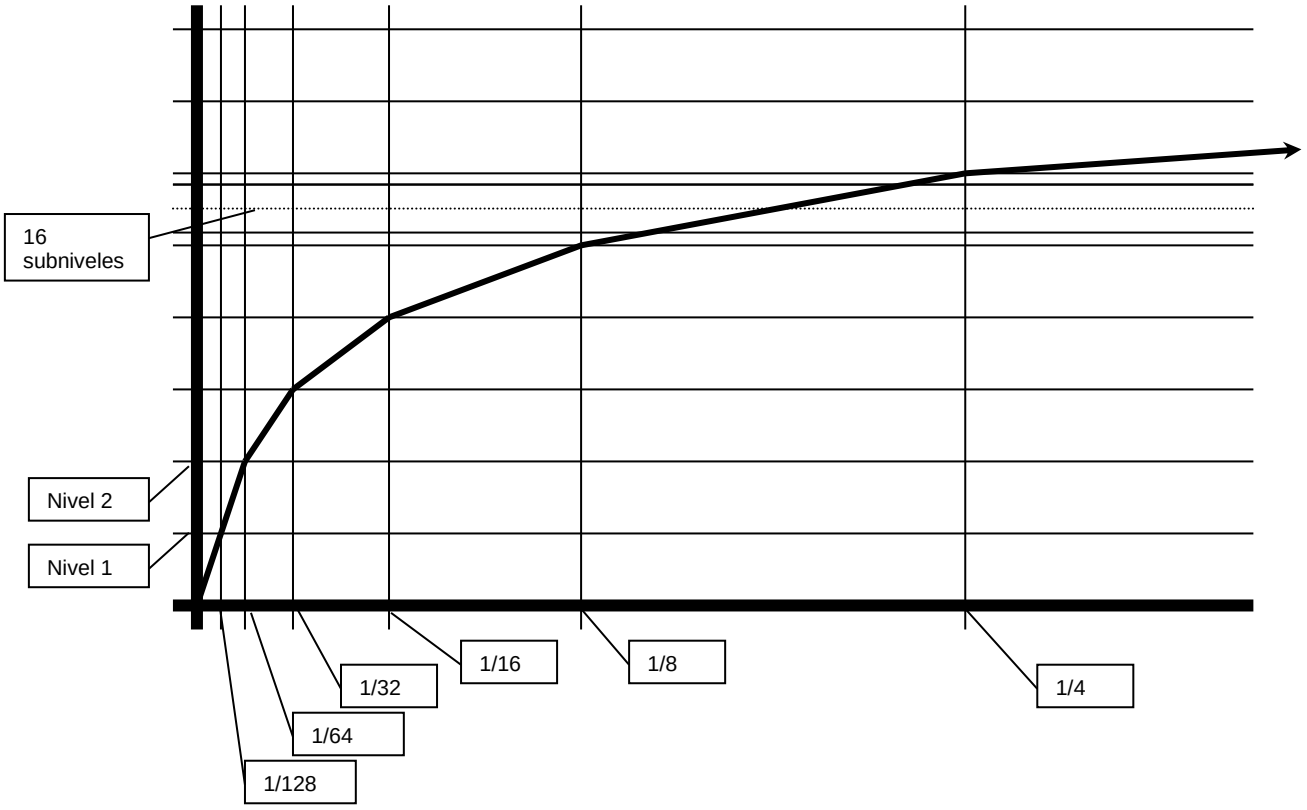
Estas "leyes de cuantificación" estandarizan en 256 niveles no lineales la cuantificación y codificación de la voz en telefonía, basadas en las fórmulas descritas. Sin embargo, la implementación real de estos algoritmos utiliza segmentos de recta en lugar de los valores resultantes de las fórmulas. La "Ley A" utiliza 13 segmentos de recta para aproximarse a la fórmula teórica, mientras que la "ley μ" utiliza 15 segmentos de recta. La gráfica de la figura anterior muestra en azul la curva "teórica" y en amarillo la "curva real", implementada con segmentos de recta. Sólo se muestra la parte correspondiente a valores positivos de señales de entrada, estandarizando en 1 la amplitud máxima. Para obtener los valores negativos, basta con simetrizar la curva respecto al origen.

Implementación de la "Ley A"

La digitalización de 13 segmentos o de la "Ley A" se realiza de la siguiente forma:

1. Se divide el eje vertical (eje "y") positivo en 8 niveles iguales, trazando rectas horizontales por cada uno de éstos niveles. Sobre el eje vertical se representarán los "Niveles cuantificados".
2. Se fija un valor unitario arbitrario sobre el eje horizontal (eje "x") positivo. Sobre el eje horizontal se representará la amplitud de la señal de entrada.
3. Se marcan sobre el eje "x" los valores 1, ½, ¼, 1/8, 1/16, 1/32, 1/64 y 1/128, trazando rectas verticales por cada uno de éstos valores.
4. Los segmentos de recta se obtienen de unir las siguientes intersecciones:
    a. (Nivel 8, 1) – (Nivel 7, ½)

b. (Nivel 7, 1/2) – (Nivel 6, 1/4)
c. (Nivel 6, 1/4) – (Nivel 5, 1/8)
d. (Nivel 5, 1/8) – (Nivel 4, 1/16)
e. (Nivel 4, 1/16) – (Nivel 3, 1/32)
f. (Nivel 3, 1/32) – (Nivel 2, 1/64)
g. (Nivel 2, 1/64) – (Nivel 1, 1/128)
h. (Nivel 1, 1/128) – (Nivel 0, 0)



5. Se simetriza respecto al origen, para obtener los valores negativos.
6. De esta manera se obtienen 16 segmentos de recta (8 para los valores positivos y 8 para los valores negativos). Sin embargo, realizando una observación más detallada, los 4 segmentos más cercanos al origen, se convierten en un solo segmento, ya que todos tienen la misma pendiente. Por esto, se llega a un total de 13 segmentos.
7. Cada “Nivel” vertical, se subdivide en 16 sub-niveles, de igual amplitud.
8. Para obtener el valor digitalizado de cada muestra de la señal:
    a. Se representa el valor analógico de la muestra de la señal sobre el eje “x”, y se traza una línea vertical hasta que corte a alguno de los segmentos.
    b. La representación de la muestra se realiza con 8 bits, de la siguiente manera:
    Bit 7: Representa el signo de la muestra
    Bit 6,5,4: Representa el “Nivel” o “Segmento” dónde cayó la muestra

Bits 3,2,1,0: Representa el “sub-nivel” o “intervalo” dentro del “Segmento” más próximo al valor de la muestra

2.1.4 Codificación

La codificación establecida en la Ley A de la recomendación G.711, establece un orden de bits como se muestra a continuación.
Se presenta cada muestra de voz con 8 bits, donde el primer bit representa el signo de la muestra, los siguientes 3 el “segmento” y los últimos 4 el “intervalo” dentro de cada segmento:

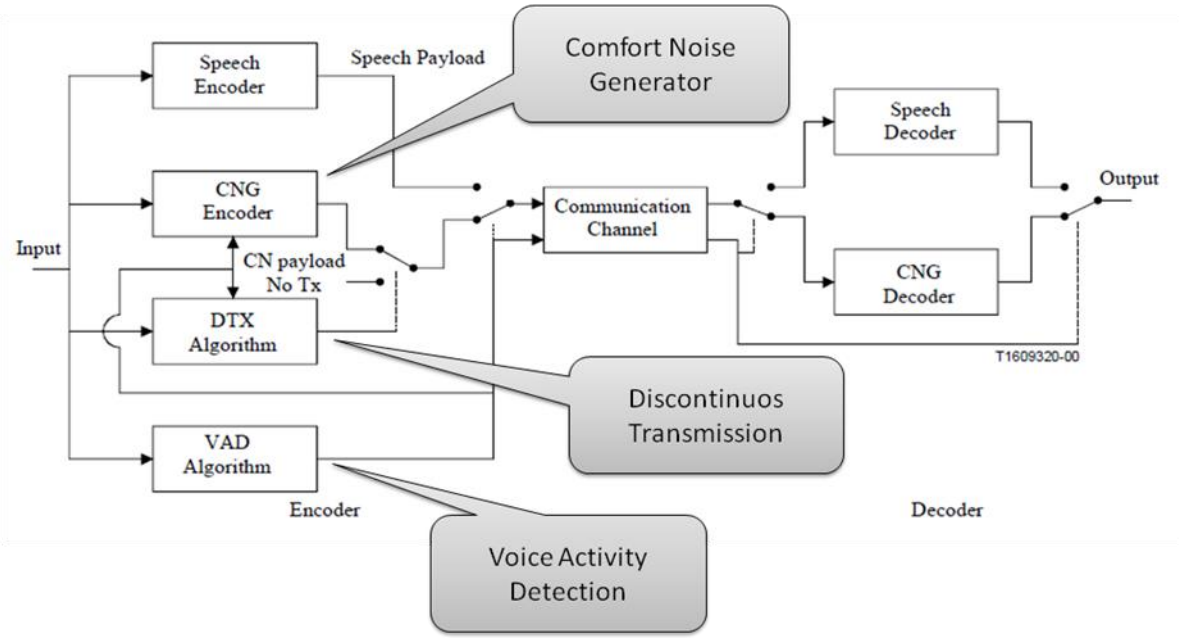
Bit	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
	Signo	**Segmento (0 - 7)**			**Intervalo (0 - 15)**			

2.1.5 G.711 Appendix II

El apéndice 2 de la recomendación ITU-T G.711 establece una definición de codificación de “ruido de confort” para ser utilizado como parte de la codificación G.711 en los sistemas de comunicación multimedia sobre redes de paquetes [17].

El formato de la codificación es genérico y puede también utilizarse con otros códecs vocales. El análisis y la síntesis de ruido de confort, así como los algoritmos de detección de actividad vocal (VAD, voice activity detection) y DTX (Discontinous Transmission) no se especifican y siguen siendo específicos de cada implementación. Sin embargo, se ha aprobado y se describe un ejemplo de solución dentro de la Recomendación

La siguiente figura esquematiza el funcionamiento mediante un diagrama de bloques. La función del algoritmo VAD es discriminar entre segmentos de voz activa e inactiva en la señal de entrada. Durante los segmentos de voz inactiva, la función del componente CNG (Comfor Noise Generator) es describir el ruido ambiente, pero reduciendo al mínimo la velocidad de transmisión (o sea, el ancho de banda necesario para su transmisión). El algoritmo DTX determina cuándo se transmite una trama SID (silence insertion descriptor). La trama SID puede enviarse periódicamente o sólo cuando hay un cambio significativo en la característica de ruido de fondo. El algoritmo CNG en el receptor utiliza la información del SID para actualizar su modelo de generación de ruido y producir luego una cantidad apropiada de ruido de confort.



.

La función del algoritmo VAD es clasificar la señal de entrada en “señal vocal activa” y “señal vocal inactiva” o un “ruido de fondo”. La clasificación incorrecta de señal vocal inactiva como señal vocal activa tiene un efecto adverso en la eficiencia del sistema, al aumentar innecesariamente la velocidad de transmisión. En este caso, la calidad vocal no es afectada. Sin embargo, cuando la señal vocal activa se clasifica indebidamente como inactiva, se recorta la señal vocal y se degrada la calidad vocal. La mayoría de los algoritmos DTX emplean un periodo de retención cuando pasan de señal vocal activa a inactiva a fin de evitar recortar el extremo final de la señal vocal. Durante el periodo de retención, las tramas de señal vocal inactiva se reclasifican como señal vocal activa.

El algoritmo DTX determina la frecuencia de la transmisión de tramas SID durante los periodos de señal vocal inactiva. Los esquemas DTX simples se actualizan periódicamente (por ejemplo, entre 5 Hz a 30 Hz). Los algoritmos DTX más complejos analizan la señal de entrada y transmiten sólo cuando se detecta un cambio significativo en el carácter del ruido ambiente.

El rol del CNG es describir y reproducir el ruido ambiente. El ruido puede describirse adecuadamente por su energía y contenido espectral. A fin de evitar cambios bruscos en el carácter del ruido de confort, es importante promediar la estimación durante un periodo de tiempo.

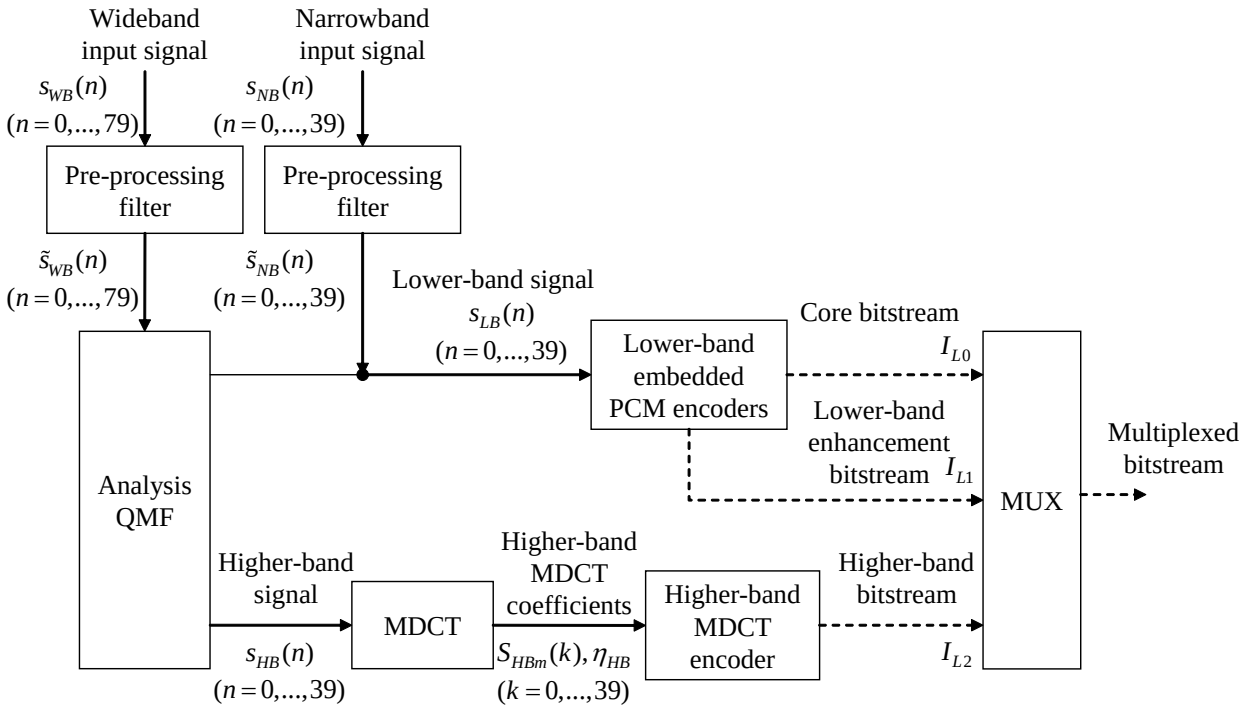
2.1.6 G.711.1

En marzo de 2008 la ITU-T aprobó un nuevo estándar de codificación de voz de banda ancha (wideband), el codec G.711.1 [12]. Esta recomendación extiende el codec G.711, el más conocido y usado en aplicaciones de telefonía, a un ancho de

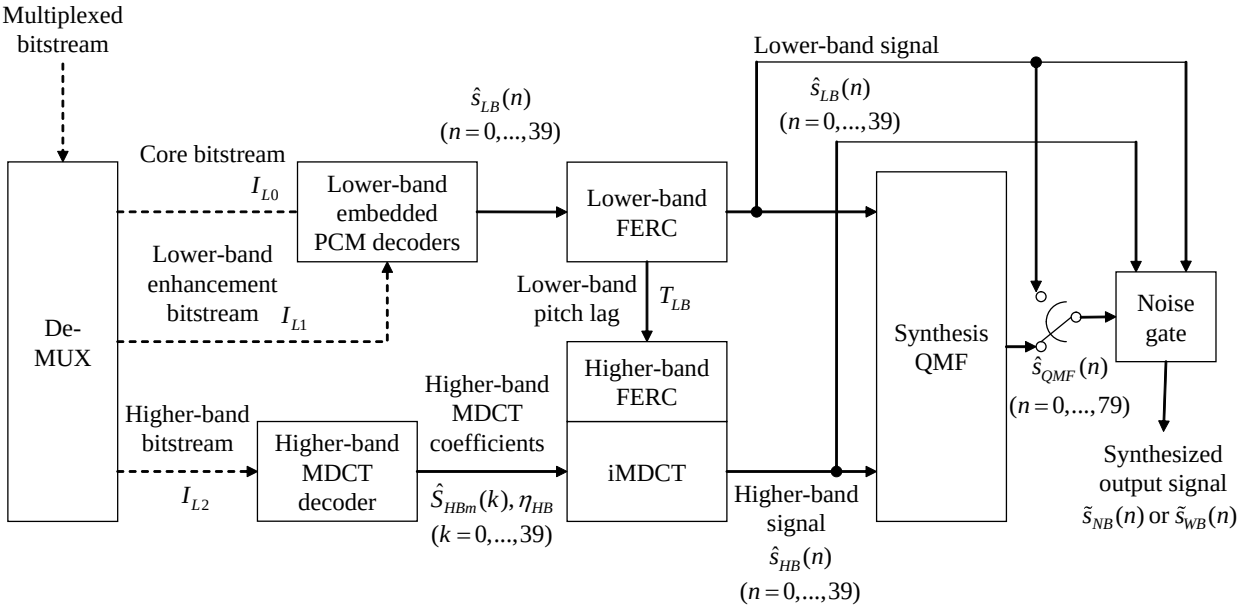
banda de 7 kHz, optimizado para aplicaciones de VoIP. Una de las características interesantes de este codec es que las muestras codificadas pueden ser convertidas en el conocido G.711 por medio de un simple truncado. El nuevo codec trabaja en 64, 80 y 96 kb/s. Las evaluaciones realizados sobre el codec muestran que cumple apropiadamente con los requerimientos establecidos [18].

Las muestras de entrada son tomadas cada 16 kHz, pero tambien está soportada la frecuencia de muestreo de 8 kHz. Tomando muestras cada 16 kHz, el ancho de banda del codec es de 50 a 7000 Hz y el resultado es de 80 o 96 kb/s. El codec utiliza tramas de 5 ms y tiene un retardo máximo de 11.875 ms.

La siguiente figura (extraída de [12]) muestra un diagrama de bloques del codificador. Un filtro pasa-altos de pre-procesamiento es aplicado a la señal de entrada muestreada a 16 kHz $s_{WB}(n)$, para filtrar los componentes menores a 50 Hz. La señal resultante, $\tilde{s}_{WB}(n)$ es dividida en una componente de baja frecuencia $s_{LB}(n)$ y otra de alta frecuencia $s_{HB}(n)$, utilizando un filtro del tipo QMF (Quadrature Mirror Filterbank). La componente de baja frecuencia es a su vez separada en las señales que entran en la banda de 300 Hz a 3.4 kHz y las que quedan fuera de esta banda. Con la primera se implementa una codificación G.711, obteniendo 64 kb/s, indicada como I_{L0} en el diagrama, o Capa 0 (“Layer 0”). La señal remanente de baja frecuencia es codificada en 16 kb/s, indicada como I_{L1} o Capa 1 (“Layer 1”).
La señal de banda alta es transformada utilizando MDCT (Modified Discrete Cosine Transform) y los coeficientes obtenidos son codificados en una señal de 16 kb/s I_{L2} o Capa 2 (“Layer 2”). Las tres capas I_{L0}, I_{L1} y I_{L2} son multiplexadas en un único flujo de salida, de 64 + 16 + 16 = 96 kb/s.



El decodificador se esquematiza en la figura siguiente (extraída de [12]). Las tramas de entrada son de-multiplexadas en la capa 0, I_{L0} compatible con G.711, la capa 1 I_{L1} de mejoras en la banda baja y la capa 2 I_{L2}, correspondiente a la banda alta del espectro de la señal de entrada. Las capas 0 y 1 son enviadas a un decodificador de banda baja. La capa 2 es enviada a un decodificador de banda alta, y luego enviada a un inversor de la transformado MDCT, marcado como iMCDT en la figura. Para mejorar la calidad ante pérdidas de tramas debido a errores en la transmisión (por ejemplo pérdida de paquetes), se implementan algoritmos de compensación, indicados como FERC (Frame ERasure Concealment). Finalmente las señalas de ambas bandas $\hat{s}_{LB}(n)$ y $\hat{s}_{HB}(n)$ son combinadas utilizando un filtro de síntesis QMF, generando la señal de banda completa $\hat{s}_{QMF}(n)$. Sobre esta señal se aplica un procesamiento de ruido, para reducir ruidos de fondo de bajo nivel, terminando finalmente en la señal de 16kHz $\hat{s}_{WB}(n)$, o de 8 kHz $\hat{s}_{NB}(n)$, según se requiera.



La codificación resultante puede operar en 4 modos, según se muestra en la tabla siguiente, generando tasas de bits de 64,80 o 96 kb/s. En el caso de la trama R2, no se provee información acerca de si contiene información de mejoras de la banda baja o de la banda alta, por lo que esto debe ser especificado en forma explícita al decoder.

Mode	**Sampling rate (kHz)**	**Core layer (Layer 0, I_{L0})**	**Lower-band enhancement layer (Layer 1, I_{L1})**	**Higher-band enhancement layer (Layer 2, I_{L2})**	**Overall bit rate (kbit/s)**
		64 kbit/s	**16 kbit/s**	**16 kbit/s**	
R1	8	x	–	–	64
R2a	8	x	x	–	80
R2b	16	x	–	x	80
R3	16	x	x	x	96

Las tramas de G.711.1 son de 5 ms y tienen 320 bits de la capa 0 (G.711), correspondientes a 8 bits x 40 muestras, 80 bits de la capa 1 y 80 bits de la capa 2, completando un total de 480 bits por trama.
La demora total del algoritmo lleva 5 ms para la información de la trama, 5 ms extras necesarios para el análisis MCDT (“lookahead”) y 1.875 ms para la implementación del filtro QMF, completando un total de 11.875 ms

2.1.7 G.729

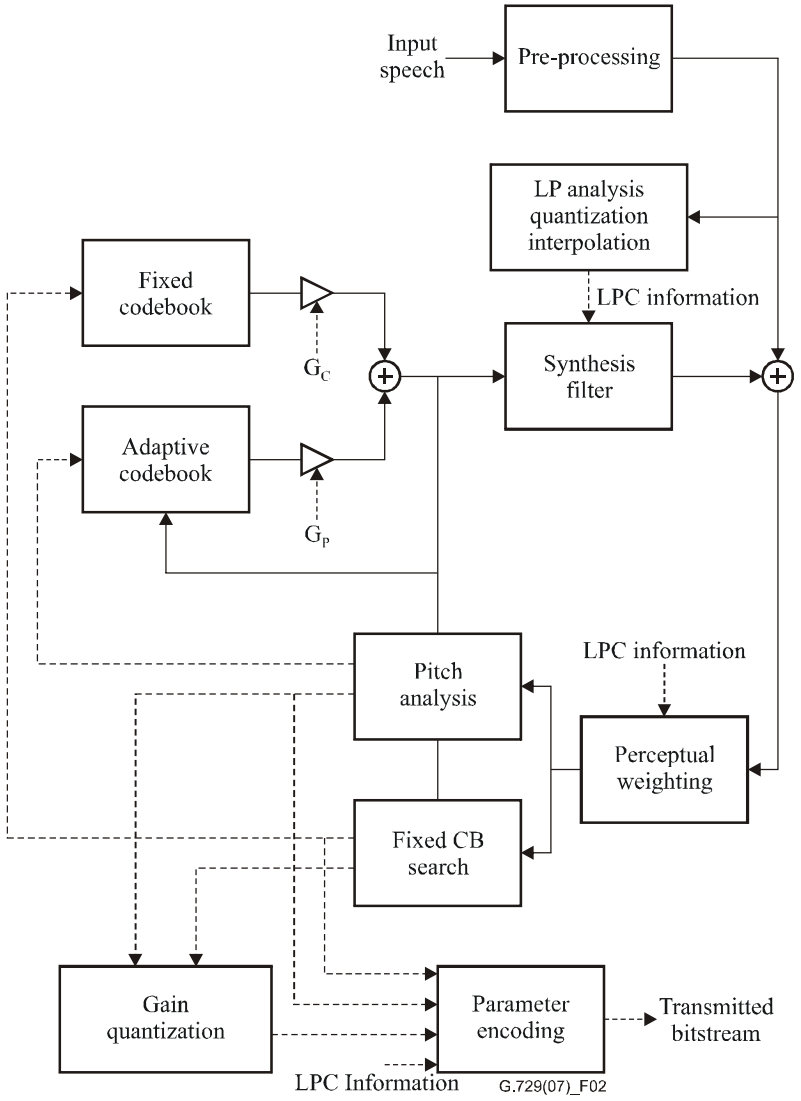
El codec G.729 [7] es un estándar de codificación para señales de audio desarrollado por la ITU, codificando las señales de voz a 8 kbit/s utilizando CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction).

Se basa en el modelo de síntesis de voz presentado en la Introducción (Sección 1). Utiliza un modelo basado en dos generadores de impulsos combinados. Estos generadores de impulsos se seleccionan de una lista predeterminada (llamada "libro de códigos" o "codebook", y se codifica el "puntero" al generador seleccionado. La técnica es conocida como CELP (Code Excited Linear Prediction), y fue propuesta en 1985 por Schroeder y Atal [19]. Utiliza ventanas de audio de 10 ms correspondientes a una cantidad de 80 muestras (ya que la frecuencia de muestreo es de 8.000 muestras por segundo). El receptor genera una nueva forma de onda que reproduce la voz en base a una síntesis basada en los parámetros codificados, como se ha descrito en la sección 1.

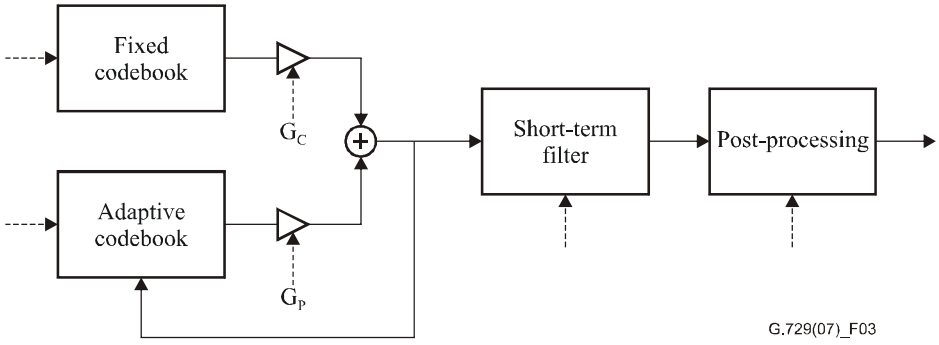
Cada 10 ms se extraen los parámetros del modelo CELP: coeficientes del filtro lineal predictivo (LPC), punteros a la tabla de impulsos adaptativos y fijos (codebooks) y ganancias). A partir de los coeficientes LPC, se obtienen parámetros equivalentes, llamados LSP (Line Spectrum Pairs) y se cuantizan usando vectores predictivos de dos etapas (VQ).

Dado que a la salida del codificador la tasa de bits es de 8 kbit/s y se toman cuadros de 10 ms, se usan 80 bits (10 bytes) para representar a cada cuadro o ventana de audio en G.729.

La siguiente figura (extraída de [7]) presenta un diagrama de bloques de un codificador G.729

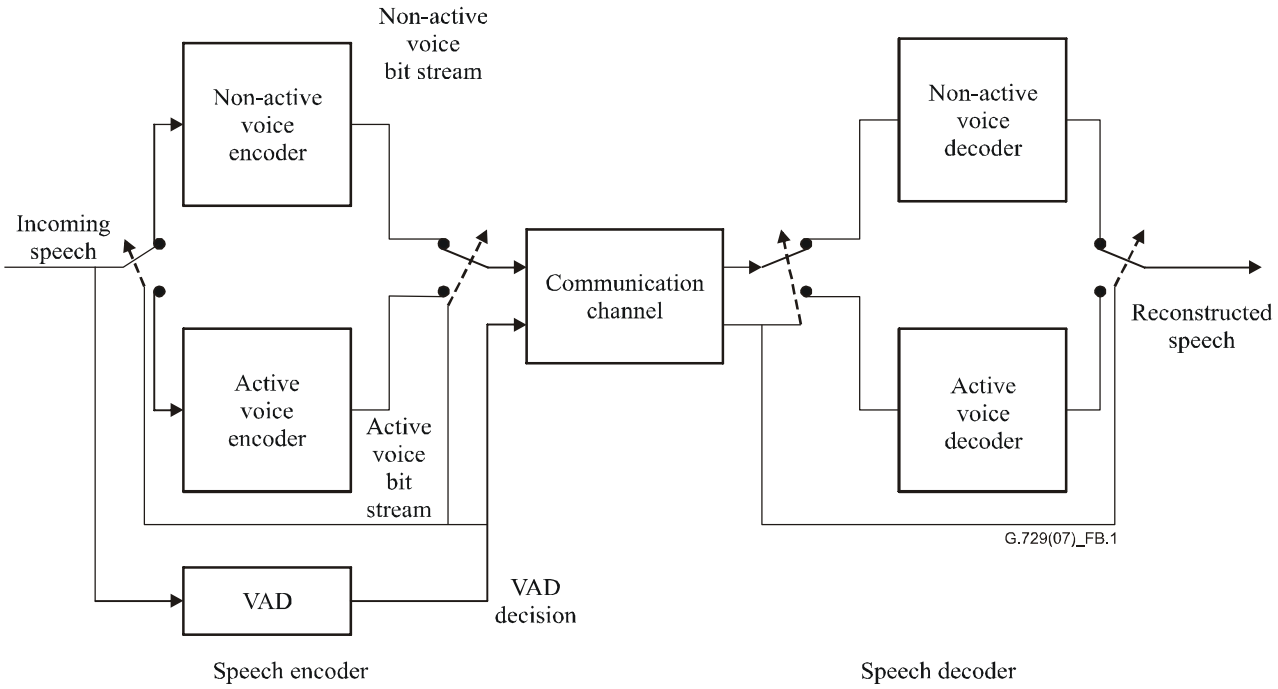


La siguiente figura (extraída de [7]) presenta un diagrama de bloques de un decodificador G.729



En el anexo A de la recomendación G.729 se define una variante de este codec logrando así un codec de menor complejidad llamado G.729A. Este es interoperable con G.729, pudiéndose utilizar un codificador G.729A con un decodificador G.729 y viceversa. Los cambios respecto a la versión original se deben a simplificaciones en los algoritmos empleados. La reducción de la complejidad incluye la sustitución de algunos bloques de procesamiento por otros más sencillos. También incluye el mantener fijos parámetros que en la versión completa del codec varían dependiendo del audio a codificar.

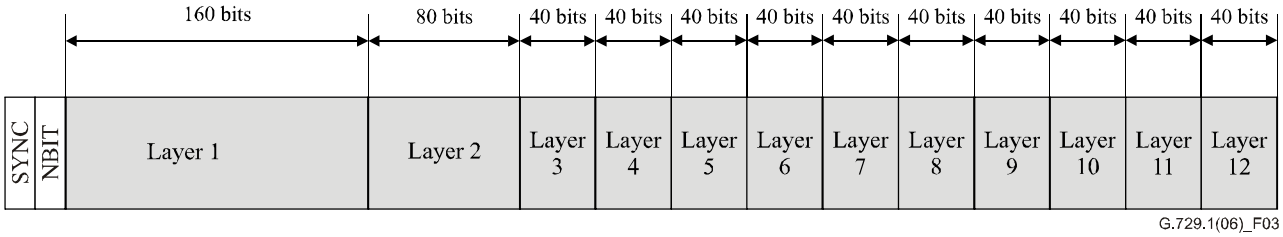
El Anexo B de la recomendación G.729 B provee detección de actividad de voz y silencios y modelado y regeneración del "ruido de fondo", lo que redunda en una disminución del ancho de banda total utilizando, ya que no se transmiten muestras durante los períodos de silencio. En la siguiente figura (extraída de [7]) se muestra esquemáticamente el proceso de detección de actividad de voz (VAD, Voice Activity Detection) y su transmisión desde el codificador al decodificador.



2.1.8 G.729.1

En mayo de 2006 la ITU-T aprobó un nuevo estándar de codificación de voz de banda ancha (wideband), el codec G.729.1. Este codec fue diseñado para proveer una transición sencilla en el mundo de la telefonía entre sistemas que utilizan banda angosta (300 a 3400 Hz) y nuevos sistemas que soporten banda ancha (50 a 7000 Hz), proveyendo interoperabilidad y compatibilidad con la recomendación G.729 y sus anexos A y B, los que tienen amplia difusión en el mundo de VoIP [20].

La señal codificada tiene una tasa de bits de 8 a 12 kb/s para señales de banda angosta (de 50 a 4000 Hz) y de 14 a 32 kb/s para señales de banda ancha (de 50 a 7000 Hz). La trama de salida consiste en 12 capas, cada una correspondiente a una tasa de bits entre los 8 y los 32 kb/s, como se muestra en la siguiente figura:



La capa 1 se corresponde con la codificación basada en CELP, es de 8kb/s y es compatible con G.729. La capa 2 se corresponde con mejoras en las frecuencias de la banda baja (50 a 4000 Hz), y ocupa 4 kb/s. Las capas siguientes agregan progresivas mejoras en la banda alta, cada una de ellas ocupando 2 kb/s adicionales.

Esta trama puede ser truncada a la salida del codificador, en el decodificador, o en cualquier punto de la red, si fuera necesario reducir el ancho de banda.

Las tramas o cuadros son de 20 ms, y la demora total del algoritmo es de 48.9375 ms, debido a los tiempos necesarios en los filtros utilizados internamente en el proceso de codificación.

2.1.9 G.723.1

El codec G.723.1 [5] es un estándar de codificación para señales de audio desarrollado por la ITU, codificando las señales de voz a 6.4 o 5.3 kbit/s. Utiliza ventanas de audio de 30 ms.

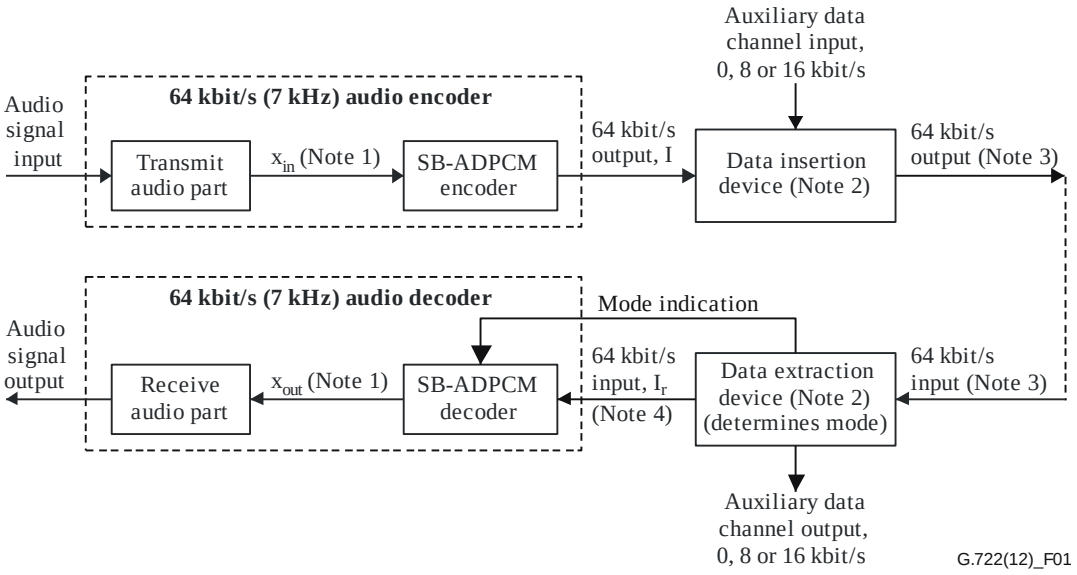
Para la codificación a 6.4 kb/s se utiliza un algoritmo MPC-MLQ (Multi-Pulse Maximum Likelihood Quantization), generando 24 bytes por cada ventana de 30 ms. Para la codificación a 5.3 kb/s se utiliza ACELP, generando 20 bytes por cada ventana de 30 ms.

El retardo total (latencia) del algoritmo es de 37.5 ms, ya que, una vez recibida la ventana de 30 ms, el algoritmo requiere de 7.5 segundos de muestras adicionales.

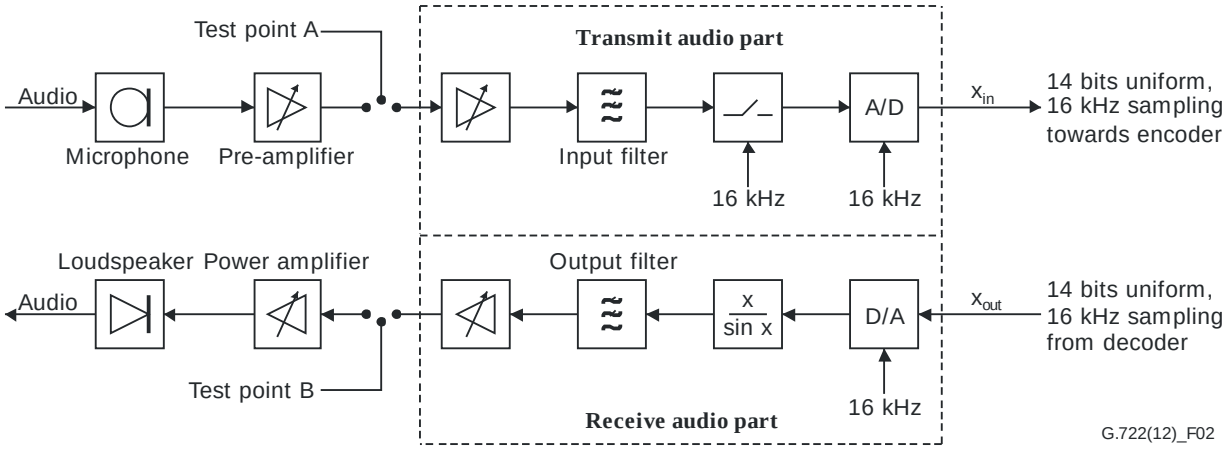
El Anexo A de la recomendación G.723.1 provee modelado y regeneración del “ruido de fondo”, lo que redunda en una disminución del ancho de banda total utilizando, ya que no se transmiten muestras durante los períodos de silencio.

2.1.10 G.722

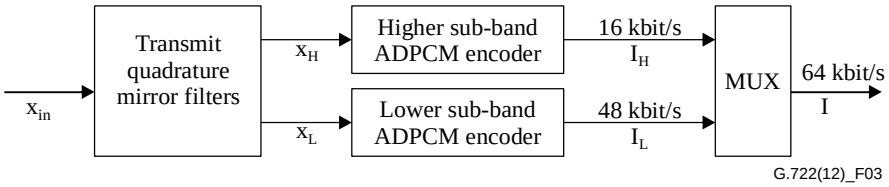
El codec G.722 es un estándar de codificación para señales de audio de banda ancha (50 Hz a 7 kHz) desarrollado por la ITU. Separa a la señal de audio en dos bandas, y cada una de ellas la codifica con técnicas ADPCM (Adaptive Differential Pulse Code Modulation). El proceso completo se identifica como SB – ADPCM (Sub Band ADPCM). Puede operar en tres modos diferente, generando bit rates de 64, 56 y 48 kbit/s a nivel de la codificación. En los últimos dos casos, junto con la codificación de audio, es posible enviar un canal auxiliar de información de 8 o 16 kbit/s, respectivamente (completando de esta manera el bit rate constante de 64 kbit/s.
La siguiente figura ilustra un diagrama de bloques de alto nivel de un codificador y un decodificador G.722:



El componente inicial es un digitalizador lineal, de 14 bits y 16 kHz, tal como se muestra en la siguiente figura

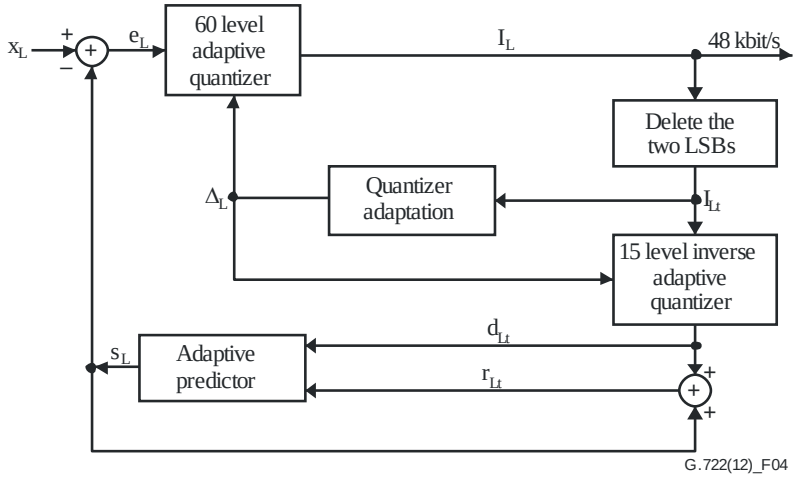


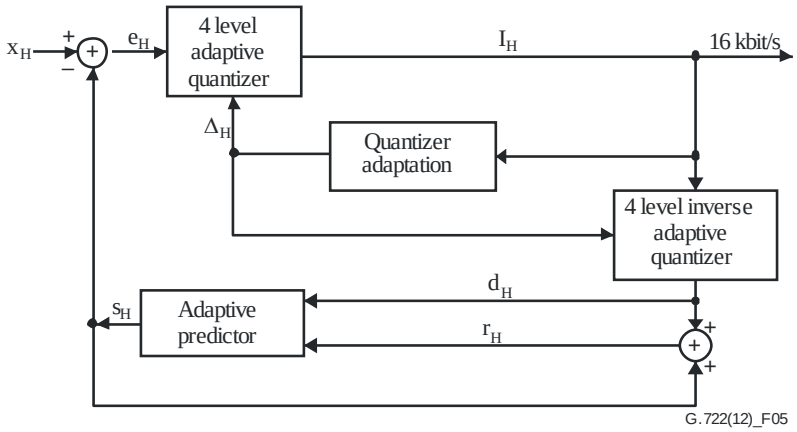
El codificador SB-ADPCM se ilustra en la siguiente figura:



Consiste en un filtro que separa en dos sub-bandas: Una correspondiente a la "banda baja" (X_L), entre 0 Hz y 4 kHz y otra a la "banda alta" (X_H), correspondiente frecuencias mayores a 4 kHz. La banda baja se codifica a 48 kbit/s y la banda alta a 16 kbit/s. Ambas son luego multiplexadas, para conformar el flujo de salida a 64 kbit/s.

Las siguientes figuras muestran los detalles de los bloques codificadores de la banda baja y la banda alta, respectivamente:





2.1.11 RTAudio

El codec RTAudio, desarrollado por Microsoft, está comenzando a ser utilizado comercial y corporativamente. Utiliza un ancho de banda de 8.8 k bit/s, con técnicas LPC (Linear Prediction Coefficients).

RTAudio utiliza técnicas de codificación VBR (Variable Bit Rate), lo que significa que no todas las ventanas o cuadros de voz se codifiquen con la misma cantidad de bytes.

El retardo total (latencia) del algoritmo es menor a 40 ms

2.1.12 AMR

El codec AMR (Adaptive Multi Rate) es utilizado típicamente en redes celulares GSM. Hace uso de tecnologías DTX (Discontinous Transmition), VAD (Voice Activity Detection) para detección de actividad vocal y CNG (Confort Noise Generation).

Provee una variedad de opciones en cuanto al ancho de banda que utiliza. Puede trabajar a las siguientes velocidades 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 y 4.75 kb/s.

De forma similar a G.729, se basa en el modelo CELP, operando con ventanas de audio de 20 ms correspondientes a una cantidad de 160 muestras (ya que la frecuencia de muestreo es de 8.000 muestras por segundo). Cada ventana de 20 ms es a su vez dividida en 4 sub-ventanas, de 5 ms (40 muestras) cada una.

Pora cada ventana se extraen los parámetros LP del modelo CELP (los coeficientes de los filtros LP), y por cada sub-ventana se obtienen los índices de los “codebooks” fijos y adaptivos y las ganancias. Estos parámetros se cuantizan y

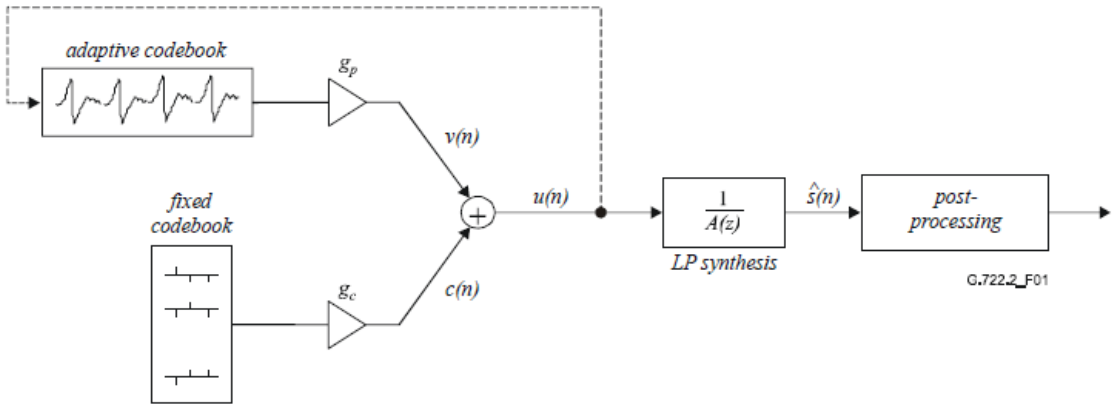
se transmiten dentro de una trama con un formato preestablecido en la recomendación del Codec.

Según la forma en que se cuanticen estos parámetros (de acuerdo a cuantos bits se utilicen para cada parámetro) se obtienen tramas de 95, 103, 118, 134, 148, 159, 204 o 244 bits, las que corresponden a velocidades de transmisión que varían entre 4.75 y 12.2 kb/s.

2.1.13 G.722.2 / AMR-WB

El codec AMR-WB (Adaptive Multi Rate – Wide Band) fue estandarizado en la Recomendación ITU-T G.722.2. Es un codec de banda ancha, de uso común con aplicaciones 3GPP (3GPP TS 26.171) y de VoIP.
Toma 14 bits por cada muestra y tiene 9 posibles velocidades de codificación, entre 6.6 y 23.85 kb/s. Está basado en las técnicas CELP, utilizando filtros de 16 polos (orden 16).

La siguiente figura, tomada de [11], esquematiza el proceso de codificación



2.1.14 SILK

SILK es el codec utilizado por Skype. Utiliza un ancho de banda variable, entre 6 a 40 kb/s, trabajando entre las bandas angostas (narrow band), con frecuencias de muestreo de 8 kHz y las bandas super anchas (superwideband), con frecuencias de muestreo de 24 kHz.

Utiliza tramas de 20 ms y tiene un retardo de 25 ms.

Desde marzo de 2009 las licencias de uso de SILK son gratuitas. En marzo de 2010 el codec fue enviado como borrador de RFC al IETF.

SILK fue reemplazado por el codec OPUS, el que finalmente fue aceptado con el RFC 6716 en setiembre de 2012

2.1.15 OPUS

OPUS es la evolución de SILK y puede trabajar tanto con CBR (Constant Bit Rate) como en VBR (Variable Bit Rate). Puede operar en diferentes bandas, llegando a ser un códec del tipo “full band”.

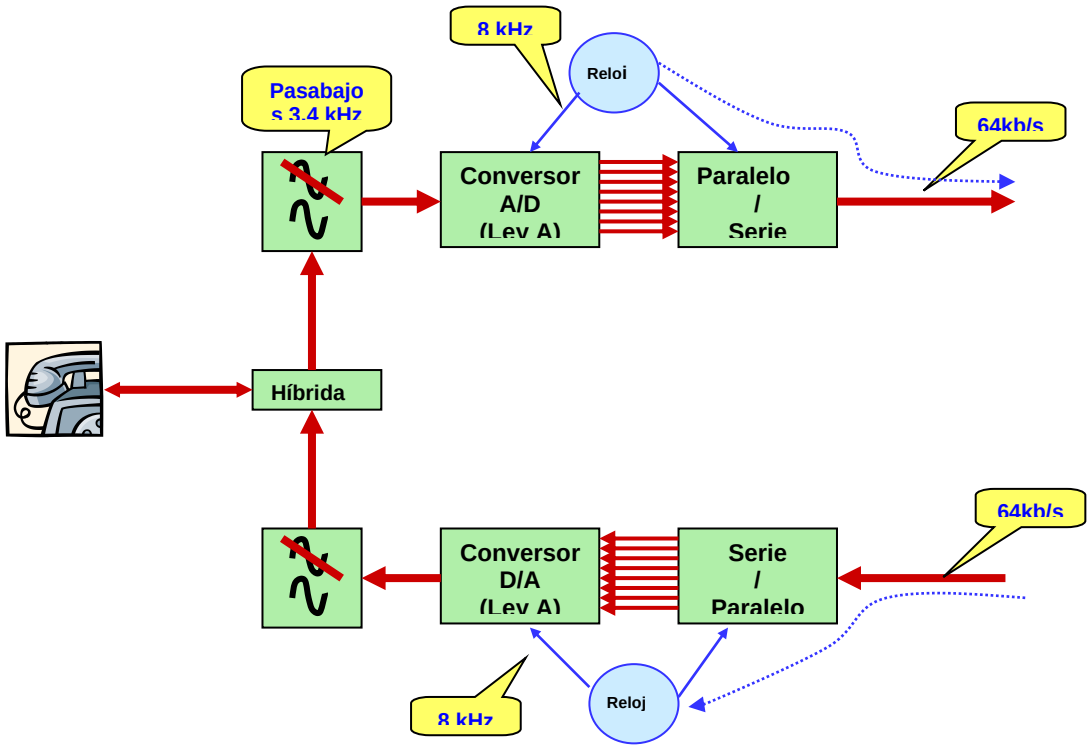
	Ancho de banda del audio	**Bit rate (kb/s)**
NB (Narrowband)	4 kHz	8 – 12 kb/s
WB (Wide Band)	8 kHz	16 – 20 kb/s
FB (Full Band)	20 kHz	28 – 40 kb/s para voz 48 - 64 kb/s para música “mono” 64 – 128 kb/s para música estereo

Es posible utilizar “ventanas” de 2.5, 5, 10, 20, 40, o 60 ms. Típicamente se utiliza 20 ms. Permite combinar múltiples ventanas en paquetes de hasta 120 ms.

2.2 Proceso de digitalización de voz en telefonía

Cómo se mencionó anteriormente, el proceso de digitalización puede realizarse en los propios teléfonos (cómo es el caso en los “teléfonos digitales” o en los “teléfonos IP”), en “Gateways” (o conversores de medios y señalización) o las “interfaces de equipo periférico” en las PBX o las “placas de abonado” en las centrales públicas.

En la siguiente figura se ejemplifica el proceso de digitalización, para el codec G.711, y a continuación se describen los componentes principales del proceso.



Híbrida
Este dispositivo es el encargado de convertir la señal analógica de “2 a 4 hilos”. Separa por canales diferentes el audio entrante del audio saliente, para que sea posible el proceso de digitalización

Pasabajos
El sistema de digitalización requiere acotar el ancho de banda de la señal de entrada a 3.4 kHz, de manera de asegurar el cumplimiento del “Teorema del Muestreo”.

Conversores A/D y D/A
Implementan la conversión digital analógica y analógica digital, con “Ley A” o “Ley μ”. Para ello se basan en un reloj de 8 kHz. Por cada muestra se obtienen 8 bits, los que son serializados.

Paralelo/Serie – Serie/Paralelo
Este proceso obtiene los 8 bits de cada muestra, y los “serializa”. De esta manera se obtiene un flujo de 8 bits x 8 kHz = **64 kbits/s,** velocidad de transmisión básica en Telefonía.

3 Introducción a la codificación de video

Los estudios acerca de la codificación de imágenes y video comenzaron en la década de 1950. En 1984 fue introducida la estrategia de codificación utilizando la transformada discreta de coseno (DCT) [21], técnica ampliamente utilizada en los sistemas actuales de codificación. Las técnicas de compensación de movimiento aparecieron también en la década de 1980, dando origen a las tecnologías híbridas MC/DCT (Motion Compensation/Discrete Cosine Transform), utilizadas en los actuales algoritmos MPEG.

Por otra parte, las transformadas discretas de Wavelets (DWT) comenzaron también a ser utilizadas en codificación de imágenes en la década de 1980, y fueron adoptadas más recientemente dentro de las tecnologías MPEG-4 y JPEG 2000, para la codificación de imágenes fijas.

La complejidad de codificadores y decodificadores ha ido aumentando, logrando un muy alto nivel de compresión, a expensas de requerir decodificadores y, sobre todo, codificadores muy complejos, y que requieren gran capacidad de procesamiento [22]. Es de esperar que en el futuro próximo se requiera aún mayor capacidad de procesamiento, reduciendo los requerimientos de ancho de banda y mejorando la calidad percibida.

Las técnicas utilizadas para la digitalización del video incluyen los siguientes conceptos:

- **Predicción**

  Mediante este proceso, se trata de “predecir” el valor de ciertas muestras en función de otras, de manera de poder enviar únicamente como información la diferencia, la que típicamente requiere menor ancho de banda para ser transmitida. En el receptor, la misma predicción es realizada, y se le aplica la diferencia (o el valor residual) que es enviado por el codificador.
  Dada la alta redundancia de información que tienen típicamente las escenas de video, esta predicción se puede realizar tanto dentro de un mismo cuadro, como entre cuadros.

- **Transformación**

  Los valores relacionados a las muestras pueden ser transformados en otro conjunto de valores equivalentes, que representan la misma información de manera diferente (por ejemplo, una misma señal puede ser representada por su amplitud en el tiempo o en el dominio de la frecuencia). En video se utiliza típicamente la “Transformada Discreta del Coseno” o DCT por sus siglas en inglés.

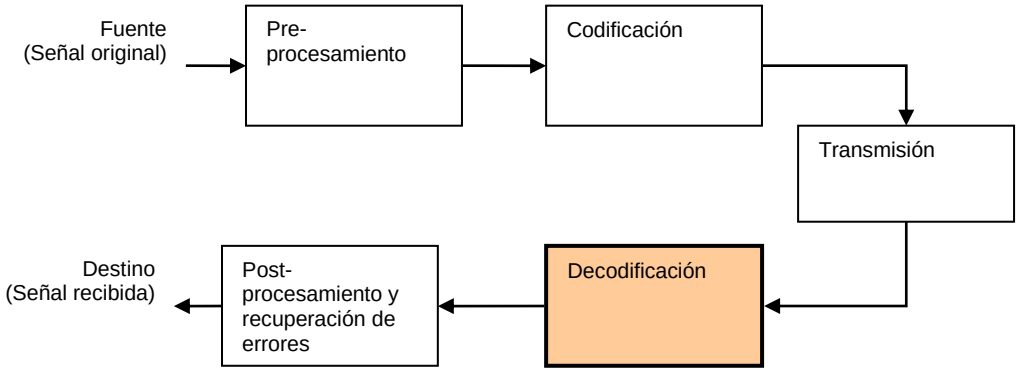
- **Cuantización**

  Es el proceso mediante el cual se asigna un valor “entero” a un número “real”. En función de la cantidad de enteros utilizados (o la cantidad de bits necesarios para su presentación), el proceso de cuantificación puede introducir más o menos distorsión respecto al valor orignal.

- **Codificación entrópica (Entropy Coding)**

  Se trata de representar los valores cuantizados de manera de tomar ventaja de las frecuencias relativas con las que aparece cada símbolo. Uno de los conocidos mecanismos de codificación entrópica es utilizar códigos de largo variable (o “VLC” por sus siglas en inglés), de manera de asignarlo a los valores que se repiten con mayor frecuencia los códigos de menor longitud.

Una cadena típica de codificación, transmisión y decodificación de video se muestra en la siguiente figura. Las estandarizaciones de ITU-T y ISO/IEC JCT se centran en detallar el proceso de decodificación, resaltado en la figura. Estas recomendaciones establecen sintaxis específicas del flujo de información. El objetivo es que cualquier decodificador que cumpla con la recomendación apropiada, pueda reproducir un flujo de video apropiadamente codificado.

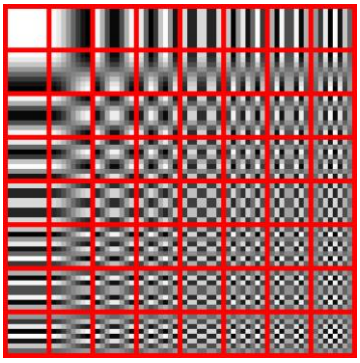


A continuación se presentan, en forma resumida, las características más destacables de las tecnologías actuales en codificación de imágenes y video, y la manera de codificar video para su transmisión sobre redes IP. No es el objetivo principal de este documento presentar un detalle pormenorizado de estas tecnologías, por lo que sólo se describirán brevemente sus características más resaltables.

4 Digitalización y codificación de video

4.1 JPEG

JPEG (Joint Photographic Experts Group) [23] es un estándar diseñado para comprimir imágenes fijas, tanto en color como en blanco y negro. El objetivo principal de este estándar fue el de lograr compresiones adecuadas, optimizando el tamaño final de los archivos comprimidos, admitiendo pérdida de calidad en la imagen. El algoritmo utilizado divide a la imagen en bloques de 8 x 8 píxeles, los que son procesados en forma independiente. Dentro de cada uno de estos bloques, se aplica la transformada discreta de coseno (DCT) bidimensional, generando para cada bloque, una matriz de 8 x 8 coeficientes. La gran ventaja de estos coeficientes, es que decrecen rápidamente en valor absoluto, lo que permite despreciar gran parte de ellos (ya que representan información de alta frecuencia espacial).
Conceptualmente, puede considerarse que cada bloque de 8 x 8 está compuesto por una suma ponderada de 64 tipos de bloques base, como se muestran en la siguiente figura. En esta figura, cada bloque corresponde con un patrón determinado. El primer bloque (arriba a la izquierda) no tiene textura. El coeficiente asociado a este bloque se corresponde con la componente de luminancia promedio del bloque. Es conocido también como componente de DC, haciendo analogía con la “componente de continua” de una señal eléctrica. El resto de los bloques presentan patrones bien definidos, con frecuencias espaciales crecientes hacia la parte inferior-derecha de la figura.

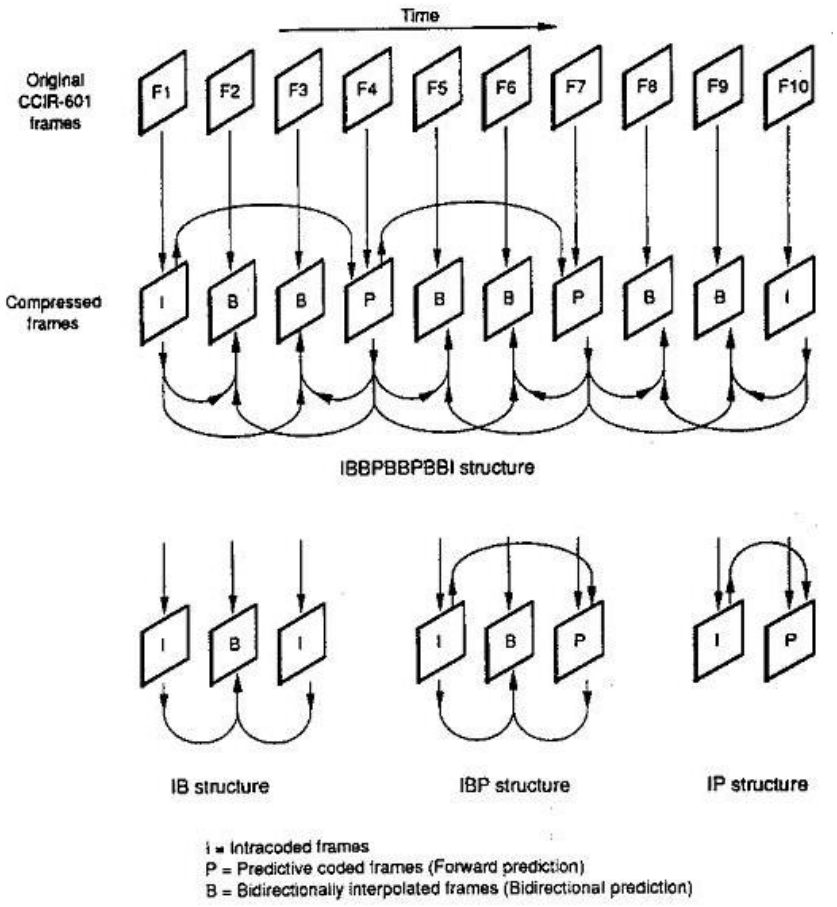


El estándar JPEG 2000 [24] está también basado en la idea de utilizar para la codificación los coeficientes de una transformación, pero en este caso se utilizan transformadas discretas de Wavelets (DWT). Esta transformada permite comprimir aún más las imágenes que la DCT. Una de las principales diferencias entre JPEG y JPEG2000 es que en esta última no es necesario dividir la imagen original en bloques. La transformada DWT se aplica a toda la imagen, lo que elimina el conocido “efecto de bloques”.

4.2 MPEG-x

MPEG-1 [25] fue originalmente diseñado por el “Moving Picture Experts Group” (MPEG) de la ISO (International Standards Organization) para el almacenamiento y reproducción digital de aplicaciones multimedia desde dispositivos CD-ROM, hasta velocidades de 1.5 Mb/s. MPEG-2 [26] fue el sucesor de MPEG-1, pensado para proveer calidad de video desde la obtenida con NTSC/PAL y hasta HDTV, con velocidades de hasta 19 Mb/s.

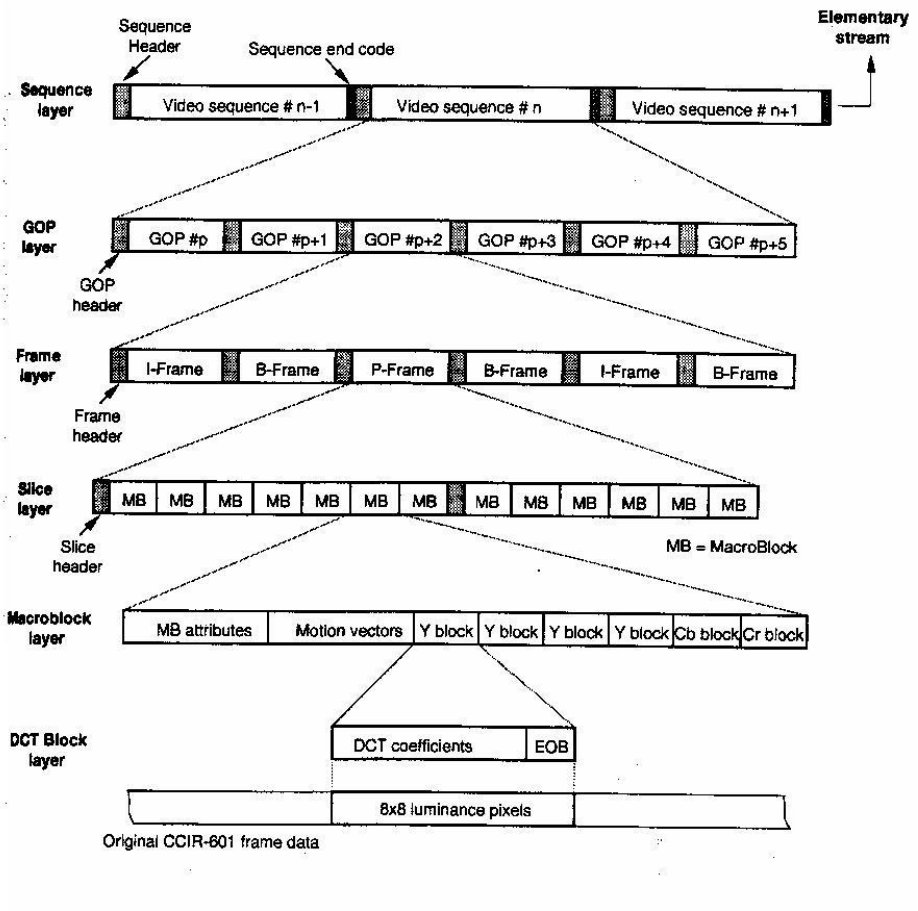
La codificación en MPEG-1 está basada en la transformada DCT para explotar las redundancias espaciales dentro de cada cuadro, y en técnicas de estimación y compensación de movimiento para explotar las redundancias temporales (entre cuadros). Las secuencias de video son primeramente divididas en “grupos de figuras” (GOP – Group of Pictures). Cada GOP puede incluir tres grupos diferentes de cuadros: I (“Intra”), P (“Predictivos”) y B (“predictivos Bidireccionales”). Los cuadros del tipo I son codificados únicamente con técnicas de compresión espacial (transformada DCT dentro del propio cuadro, por ejemplo). Son utilizados como cuadros de referencia para las predicciones (hacia adelante o hacia atrás) de cuadros P o B. Los cuadros del tipo P son codificados utilizando información previa de cuadros I u otros cuadros P, en base a estimaciones y compensaciones de movimiento. Los cuadros B se predicen en base a información de cuadros anteriores (pasados) y también posteriores (futuros). El tamaño de un GOP está dado por la cantidad de cuadros existentes entre dos cuadros I. Típicamente se utilizan de 12 a 15 cuadros para un GOP, y hasta 3 cuadros entre un I y un P o entre dos P consecutivos (típicamente una señal PAL se codifica con un GOP de tamaño 12 y una NTSC con 15, ambas con no más de 2 cuadros B consecutivos). Un ejemplo tomado de [27] se muestra en la figura (IBBPBBPBBI), donde las flechas indican los cuadros utilizados para las predicciones. Cuando más grande el GOP, mayor compresión se puede obtener, pero a su vez existe menor inmunidad a la propagación de errores.



Al igual que en JPEG, en MPEG-1 se divide la imagen de cada cuadro en bloques de 8 x 8 píxeles, los que son procesados en forma independiente. Dentro de cada uno de estos bloques, se aplica la transformada discreta de coseno (DCT) bidimensional, generando para cada bloque, una matriz de 8 x 8 coeficientes. A su vez, cuatro bloques se agrupan en un “macro-bloque” de 16 x 16 píxeles, el que es utilizado como base para la estimación del movimiento. La estimación de movimiento de un macro-bloque se realiza en el codificador, comparando el macro-bloque de una imagen con todos las posibles secciones de tamaño igual al macro-bloque (dentro de un rango espacial de 512 píxeles en cada dirección) de la(s) imagen(es) siguiente(s). La comparación se realiza generalmente buscando la mínima diferencia (el mínimo valor del error cuadrático medio MSE) entre el macro-bloque y la sección evaluada. Este procedimiento se basa en la hipótesis que todos los píxeles del macro-bloque tendrán por lo general un mismo desplazamiento, y por lo tanto, será más eficiente codificar un “vector de movimiento” del macro-bloque y las diferencias del macro-bloque predicho respecto del macro-bloque original. Las diferencias entre el macro-bloque predicho y el real también son transformadas mediante la DCT para su codificación.

Un flujo de video de MPEG-2 se forma de la manera descrita a continuación. Se utiliza como unidad básica un macro-bloque, compuesto típicamente por 4 bloques

de luminancia y 2 de crominancia (ya que la crominancia es sub-muestreada). Los coeficientes DCT de cada uno de estos bloques son serializados, y precedidos por un cabezal de macro-bloque. Varios macrobloques contiguos (en la misma fila, y de izquierda a derecha) son agrupados formando un "slice", el que a su vez es precedido de un cabezal de "slice", el que contiene la ubicación del "slice" en la imagen y el factor de cuantización usado. Típicamente puede haber un "slice" por cada fila de macro-bloques, pero puede también haber slices con parte de una fila. Un grupo de "slices" forma un cuadro, el que es precedido por un cabezal de cuadro, conteniendo información del mismo, como por ejemplo el tipo de cuadro (I,P,B), y las matrices de cuantización utilizadas. Varios cuadros se juntan, formando el GOP, también precedido de un cabezal de GOP. Finalmente, varios GOPs pueden serializarse en una secuencia (Elementary Stream), con su correspondiente cabezal, el que contiene información general, como el tamaño de los cuadros, y la frecuencia de cuadros. En la siguiente figura (tomada de [27]) se muestra un esquema del sistema de capas descrito.



MPEG-4 [28] es la evolución de MPEG-1 y 2, y provee la tecnología necesaria para la codificación en base a contenidos, y su almacenamiento, transmisión y manipulación. Presenta mejoras interesantes respecto a la eficiencia de la codificación, robustez de transmisión e interactividad. MPGE-4 puede codificar múltiples "Objetos de video" (MVO – Multiple Video Objects), ya que sus

contenidos son representados en forma individual. El receptor puede de esta manera recibir diferentes flujos por cada objeto codificado dentro de un mismo video, correspondientes por ejemplo a diferentes “planos” (VOP – Video Object Plane) de la imagen. Cada secuencia de VOPs constituye un objeto de video (VO – Video Object) independiente, los que son multiplexados dentro de una transmisión, y demultiplexados y decodificados por el receptor.

4.3 H.264

En 2001, el grupo MPEG de ISO/IEC y el VCEG (Video Coding Expert Group) del ITU-T decidieron unir esfuerzos en un emprendimiento conjunto para estandarizar un nuevo codificador de video, mejor que los anteriores, especialmente para anchos de banda o capacidad de almacenamiento reducidos [29]. El grupo se llamó JVT (Joint Video Team), y culminó con la estandarización de la recomendación H.264/MPEG-4 Part 10, también conocida como JVT/H.26L/AVC (Advanced Video Coding) o H.264/AVC en 2003. Este nuevo estándar utiliza compensaciones de movimiento más flexibles, permitiendo dividir los macro-bloques en diversas áreas rectangulares, y utilizar desplazamientos de hasta un cuarto de píxel. Agrega además los cuadros del tipo SP (Switching P) y SI (Switching I), similares a los P e I, pero con la posibilidad de reconstruir algunos valores específicos de forma exacta.
Las técnicas de codificación entrópica que utiliza utiliza H.264 son las conocidas como Context-Adaptive Variable-Length Coding (CAVLC) y Context-Adaptive Binary Arithmetic Coding (CABAC). Esta última (CABAC) es más compleja que la primera (CAVLC), pero a su vez, más eficiente.

Con H.264/AVC, para una misma calidad de video, se logran mejoras en el ancho de banda requerido de aproximadamente un 50% respecto estándares anteriores [30] [31].

En 2007 fue aprobada una extensión de H.264/AVC incluyendo el “Anexo G”, llamada “Scalable Video Coding” o SVC por sus iniciales. Esta modificación permite la construcción de sub-flujos de datos dentro de un flujo principal. El flujo principal o “capa base” (base layer) puede ser decodificado por cualquier equipo que soporte H.264/AVC, aunque no soporte SVC. Los flujos adicionales pueden contener información adicional del flujo, brindando mayor definición.

En 2010 fue agregado el “Anexo H”, llamado “Multiview Video Coding” o MVC por sus iniciales. Este agregado está pensado para permitir tener diferentes flujos representando diferentes visiones de la misma escena, y fue desarrollado para aplicaciones de video en 3D, donde son necesarios dos flujos de información para generar el efecto “estereoscópico” de una misma escena.

La recomendación H.264 establece diferentes “perfiles” y “niveles”. Los “perfiles” establecen requerimientos mínimos a cumplir por el codificador y decodificador. Se establecen 12 perfiles para el estándar base AVC, 3 para el SVC y 2 para el

MVC, completando un total de 17 perfiles según la versión 2010 de la recomendación [32]. A continuación se enumeran los perfiles establecidos en la recomendación H.264:

- Constrained Baseline Profile (CBP)
  Diseñado para aplicaciones de bajo costo. Utilizado típicamente en aplicaciones móviles y algunos servicios de video conferencias.

- Baseline Profile (BP)
  Similar al CBP, es utilizado primariamente por aplicaciones de video móvil y video conferencias. Incluye las mismas características que el CBP, agregando algunas funciones que le brindan mayor robustez frente a pérdidas de información. Solo admite cuadros del tipo I y P (no se admiten cuadros del tipo B)

- Main Profile (MP)
  Pensado para aplicaciones de TV en definición estándar (SDTV). Admite cuadros del tipo I, P y B.

- Extended Profile (XP)
  Diseñado para “streaming” de video.

- High Profile (HiP)
  Pensado para TV en alta definición (HDTV) y almacenamiento en discos (por ejemplo, es el utilizado en “Blu-ray”).

- High 10 Profile (Hi10P)
  Para aplicaciones que requieran mejor calidad que el HiP, soporta hasta 10 bits por muestra.

- High 4:2:2 Profile (Hi422P)
  Especial para aplicaciones de video “profesional”

- High 4:4:4 Predictive Profile (Hi444PP)
  Para muy alta calidad, sin compresión, soporta hasta 14 bits por muestra

- High 10 Intra Profile
  Similar al Hi10P, pero admitiendo solo cuadros I, usado en aplicaciones profesionales.

- High 4:2:2 Intra Profile
  Similar al Hi422P, pero admitiendo solo cuadros I, usado en aplicaciones profesionales.

- High 4:4:4 Intra Profile

Similar al Hi444P, pero admitiendo solo cuadros I, usado en aplicaciones profesionales.

- CAVLC 4:4:4 Intra Profile
  Similar al Hi444P, pero admitiendo solo cuadros I y codificación CAVLC.

- Scalable Baseline Profile
  Diseñado dentro de la extensión SVC, para aplicaciones de video conferencias y móviles.

- Scalable High Profile
  Diseñado dentro de la extensión SVC, para aplicaciones de broadcasting y streaming

- Scalable High Intra Profile
  Diseñado dentro de la extensión SVC, utiliza solo cuadros I

- Stereo High Profile
  Diseñado dentro de la extensión MVC, para aplicaciones de de video en 3D

- Multiview High Profile
  Diseñado dentro de la extensión MVC, soporta dos o más vistas de cada escena.

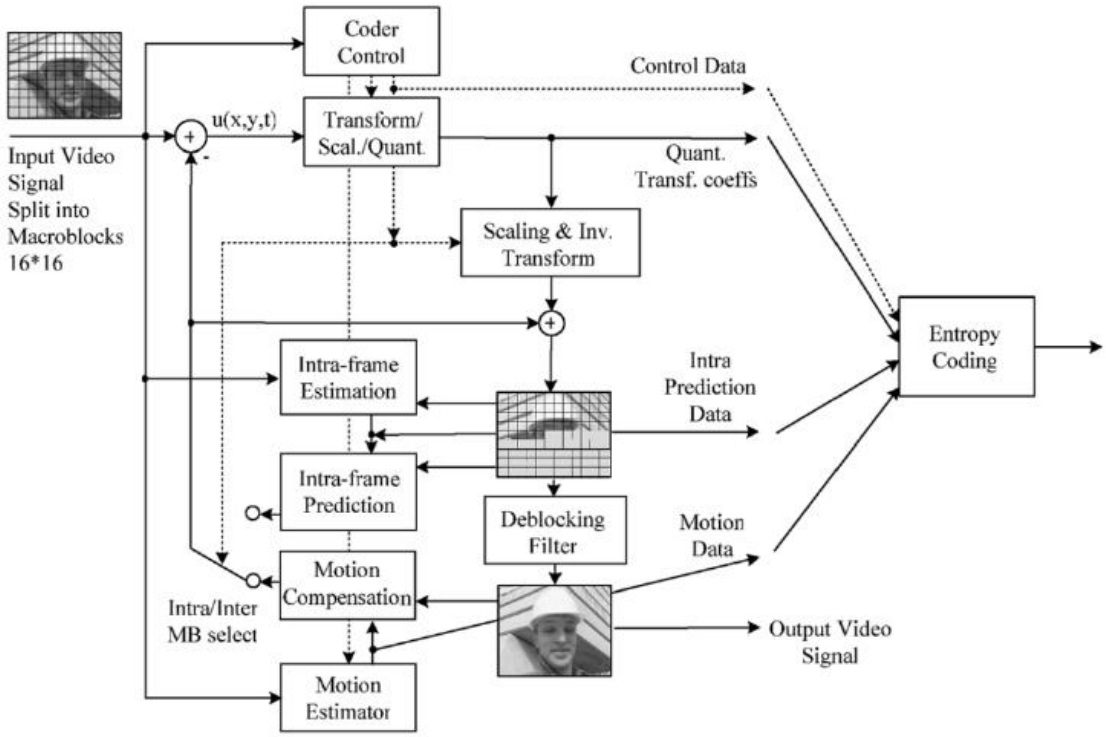
Por otra parte, los niveles establecen un conjunto de restricciones que se aplican a cada perfil. Por ejemplo, para cada perfil, un nivel puede determinar la resolución máxima de pantalla, el máximo frame rate y el máximo bit rate. Un decodificador que soporta cierto nivel, debe ser capaz de decodificar cualquier señal que tenga un nivel igual o inferior.

En la siguiente tabla se presenta un resumen comparativo de los diferentes estándares de codificación de video. Como se puede observar en dicha tabla, los codificadores / decodificadores H.264/AVC no son compatibles con los estándares anteriores, lo que supone un punto de quiebre en la evolución del video digital.

Característica	**MPEG-1**	**MPEG-2**	**MPEG-4**	**H.264/MPEG-4 Part 10/AVC**
Tamaño del macro-bloque	16x16	16x16 (frame mode) 16x8 (field mode)	16x16	16x16
Tamaño del bloque	8x8	8x 8	16x16 8x8, 16x8	8x8, 16x8, 8x16, 16x16, 4x8, 8x4, 4x4
Transformada	DCT	DCT	DCT/DWT	4x4 Integer transfor
Tamaño de la muestra para aplicar la transformada	8x8	8x8	8x8	4x4
Codificación	VLC	VLC	VLC	VLC, CAVLC, CABAC

Característica	MPEG-1	MPEG-2	MPEG-4	H.264/MPEG-4 Part 10/AVC
Estimación y compensación de movimiento	Si	Si	Si	Si, con hasta 16 MV
Perfiles	No	5 perfiles, varios niveles en cada perfil	8 perfiles, varios niveles en cada perfil	3 perfiles, varios niveles en cada perfil
Tipo de cuadros	I,P,B,D	I,P,B	I,P,B	I,P,B,SI,SP
Ancho de banda	Hasta 1.5 Mbps	2 a 15 Mbps	64 kbps a 2 Mbps	64 kbps a 150 Mbps
Complejidad del codificador	Baja	Media	Media	Alta
Compatibilidad con estándares previos	Si	Si	Si	No

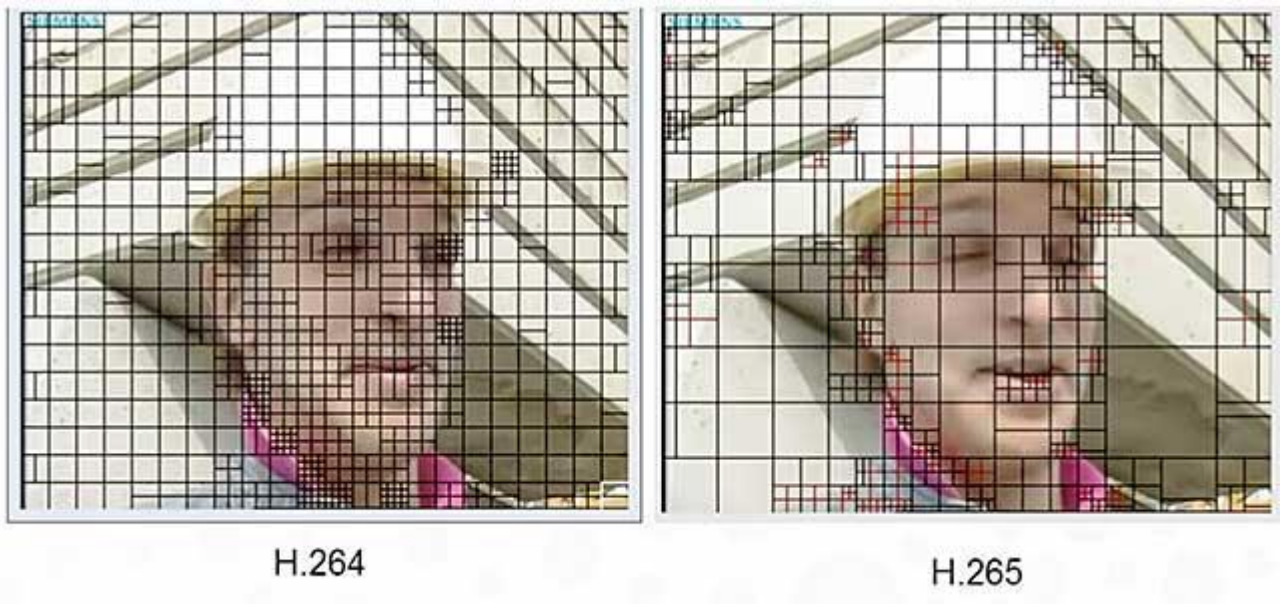
En la siguiente figura, tomada de [33], se muestra en un diagrama de bloques el proceso de codificación de video, en un codificador H.264/AVC.



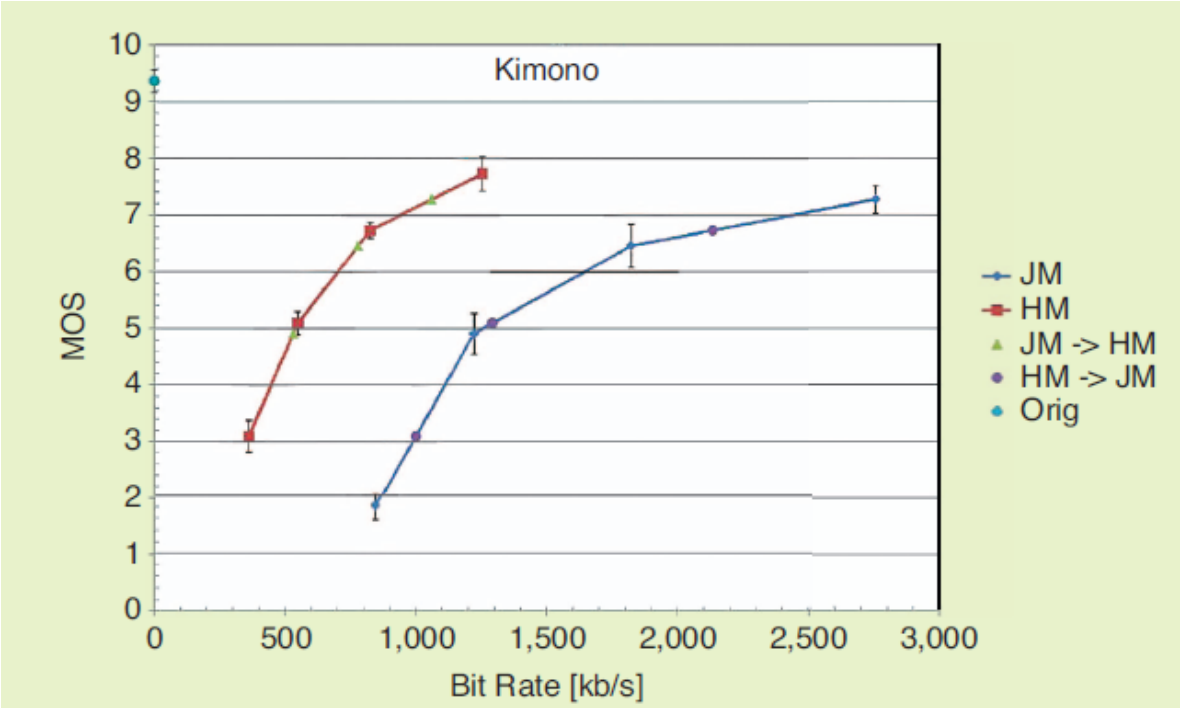
4.4 H.265

En 2013 se estandarizó la primera versión del CODEC ITU-T H.265 o MPEG-H Parte2 o High Efficiency Video Coding (HEVC) [34], desarrollado conjuntamente por la ISO/IEC Moving Picture Experts Group (MPEG) y ITU-T Video Coding Experts Group (VCEG) como ISO/IEC CD 23008-2 High Efficiency Video Coding.

HEVC reemplaza a los macrobloques, utilizados en las normas anteriores, con Unidades de Codificación en Arbol (Coding Tree Units - CTU) que pueden usar estructuras de bloques más grandes, de hasta 64 x 64 píxeles y pueden mejorar la segmentación de la imagen en estructuras de tamaño variable. Esto mejora la eficiencia de codificación, logrando mayores compresiones. Un ejemplo de los la segmentación realizada H.264 comparada con H.265 se muestra en la siguiente figura, tomada de [35].



Para una misma calidad de imagen, este nuevo códec puede reducir la tasa de bits de su antecesor, tal como se puede ver en esta gráfica [36], donde se grafica calidad versus bitrate para una misma secuencia de video codificada en H.264 y H.265



H.265 está comenzado a utilizarse en diversos dispositivos. Ha sido seleccionado como el códec preferido para las aplicaciones de video en Ultra Alta Definición (UHD) o 4k.

Referencias

[1] The History of Linear Prediction
Bishnu S Atal
IEEE Signal Processing Magazine, March 2006, pp 154-161

[2] ITU-T Coders for Wideband, Superwideband and Fullband Specch Communication
R. V. Cox, S. Ferraz de Campos Neto, C. Lamblin, M. Hashem Sherif
IEEE Communications Magazine, October 2009, Vol. 47, No. 10

[3] ITU-T G Series: Transmission systems and media, digital systems and networks
http://www.itu.int/net/itu-t/sigdb/speaudio/Gseries.htm

[4] Recommendation G.711: “Pulse Code Modulation (PCM) of voice frequencies”
CCITT, 1988.

[5] Recommendation G.723.1: “Dual Rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s”
ITU-T, May 2006.

[6] Recommendation G.728: “Coding of speech at 16 kbit/s using Low-delay code excited linear prediction”
CCITT, 1992.

[7] Recommendation G.729: “Coding of speech at 8 kbits using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)”
ITU-T, Jan 2007.

[8] Adaptive Multi-Rate (AMR) speech codec
ETSI TS 126 090 V9.0.0, 2010-01

[9] Recommendation G.722: “7 kHz audio-coding within 64 kbit/s”
CCITT, 1988.

[10] Recommendation G.722.1: “Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss”
ITU-T, 05/2005.

[11] Recommendation G.722.2: “Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)”
ITU-T, 07/2003.

[12] Recommendation G.711.1: “Wideband embedded extension for G.711 pulse code modulation”
ITU-T, 03/2008.

[13] Recommendation G.729.1: “G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729”
ITU-T, 05/2006.

[14] Overview of the Microsoft RTAudio Speech Codec
Microsoft, 2006.

[15] SILK – Super Wideband Audio Codec

https://developer.skype.com/silk

[16] Recommendation G.719: “Low-complexity, full-band audio coding for high-quality, conversational applications”
ITU-T, 06/2008.

[17] Recommendation G.711 Appendix II: “A Comfort Noise Payload Definition for ITU-T G.711 Use in Packet-Based Multimedia Communications Systems”
ITU-T, 02/2002.

[18] ITU-T G.711.1: Extending G.711 to Higher-Quality Wideband Speech
Y Hiwasaki, H. Ohmuro, NTT Corporation.
IEEE Communications Magazine, October 2009, Vol. 47, No. 10

[19] Code-excited linear prediction(CELP): High-quality speech at very low bit rates
Schroeder, M. Atal, B.
Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.
April 1985, Volume: 10, On page(s): 937- 940

[20] ITU-T G.729.1: Scalable Codec for New Wideband Services
I Varga, S Proust, H Taddei.
IEEE Communications Magazine, October 2009, Vol. 47, No. 10

[21] Discrete Cosine Transform
N Ahmed, T Natrajan, K.R. Rao
IEEE Trans. Comput. Vol C-23, No 1, pp90-93, Dec 1984

[22] Trends and Perspectives in Image and Video Coding
T Sikora
IEEE Proceedings, Vol 93, No 1, January 2005

[23] ISO/IEC IS 10918-1, ITU-T Recommendation T.81 Digital compression and coding of continuous-tone still images: Requirements and guidelines, 1994

[24] ISO/IEC 15444-1:2004. JPEG2000 Image Coding System: Core coding system

[25] ISO/IEC 11172-2:1993. Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s

[26] ISO/IEC 13818-2:2000. Information technology – generic coding of moving pictures and associated audio information: Video.

[27] Digital television fundamentals: design and installation of video and audio systems
Michel Robin, Michel Poulin
ISBN 0-07-053168-4, 1998, McGraw-Hill

[28] ISO/IEC 14496-2:2001. Information technology – Coding of audio-visual objects – Part 2: Visual

[29] The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extension
Gary J. Sullivan, Pankaj Topiwala, and Ajay Luthra
SPIE Conference on Applications of Digital Image Processing XXVII
Special Session on Advances in the New Emerging Standard: H.264/AVC, August, 2004

[30] Overview of the H.264 / AVC Video Coding Standard
Thomas Wiegand, Gary J. Sullivan, Gisle Bjontegaard, and Ajay Luthra
IEEE Transactions on Circuits and Systems For Video Technology, Vol 13, July 2003

[31] Report of The Formal Verification Tests on AVC (ISO/IEC 14496-10 | ITU-T Rec. H.264)
ISO/IEC JTC1/SC29/WG11, MPEG2003/N6231
December 2003

[32] ITU-T Recommendation H.264: Advanced video coding for generic audiovisual services
March 2010

[33] Video Compression – From Concepts to the H.264/AVC Standard
Gary J. Sullivan, Thomas Wiegand
Proceedings of the IEEEIssue 1, pp. 18 - 31, Jan 2005

[34] ITU-T Recommendation H.265: High efficiency video coding
December 2016

[35] What Is HEVC (H.265)?
Jan Ozer
February 2013
http://www.streamingmedia.com/Articles/Editorial/What-Is-.../What-Is-HEVC-(H.265)-87765.aspx

[36] High Efficiency Video Coding: The Next Frontier in Video Compression
Jens-Rainer Ohm and Gary J. Sullivan,
IEEE SIGNAL PROCESSING MAGAZINE, Jan 2013