Codificación de Voz y Audio

INTRODUCCIÓN

Dr. Ing. José Joskowicz

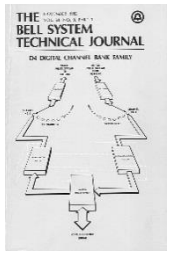
Historia

1870 – Fonógrafo
Inventado por Thomas Edison
Grabación en cilindros

1890 – Gramófono
Inventado por Emile Berliner
Grabación en discos

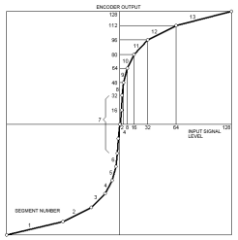
1925 – Tocadiscos eléctrico
Emile Berliner
48 RPM

1937 –PCM
Alec Reeves



1962 – Primer transmisión de audio digital T1
Bell

1963 – Cassette
Phillips



1972 – PCM “Ley A/mu”

1979 – Walkman
Sony

1981 – Compact Disc (CD)
Sony / Phillips

1993 – MP3
Estándar ISO/IEC 11172-3

1995 – Digital Versatile Disc (DVD)
IBM, Apple, Compaq, Hewlett-Packard, Microsoft, Phillips, Sony

1997 – Advanced Audio Coding (AAC)
ISOIIEC 13818-7

Historia

1999 – “MPMan”
Saehan Information Systems’s

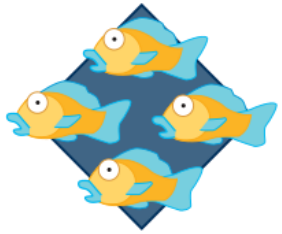
napster

1999 – Napster

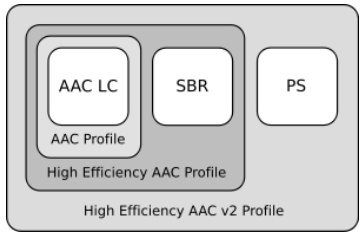
2001– iPod
Apple

Blu-ray Disc

2002 – Bue Ray



2002 – Ogg Vorbis



2004 – AAC v2

Spotify

2006 – Spotify

MPEG SURROUND

2007 – SAC o
MPEG-Surround

2014 –Enhanced Voice Services

MPEG-H AUDIO

2015 –MPEG-H 3D Audio

L Y R A

2021 –SATIN / LYRA

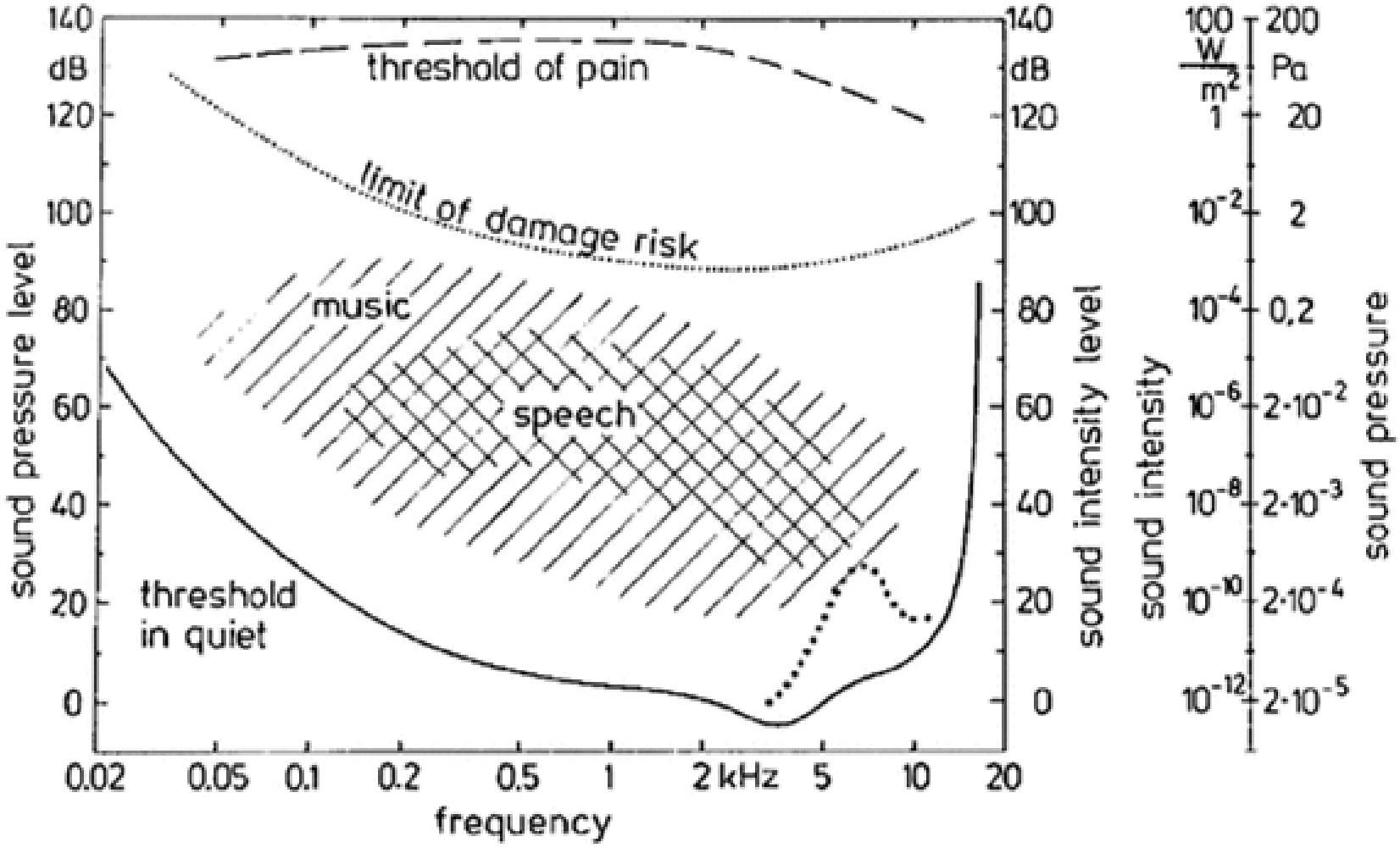
Introducción

Todos los sistemas multimedia actuales requieren digitalizar audio, es decir, convertir una señal naturalmente analógica en una secuencia de número discretos, para poder ser procesada y transmitida, y posteriormente decodificada y transformada nuevamente en analógica.

CODECS:

Codificadores / Decodificadores

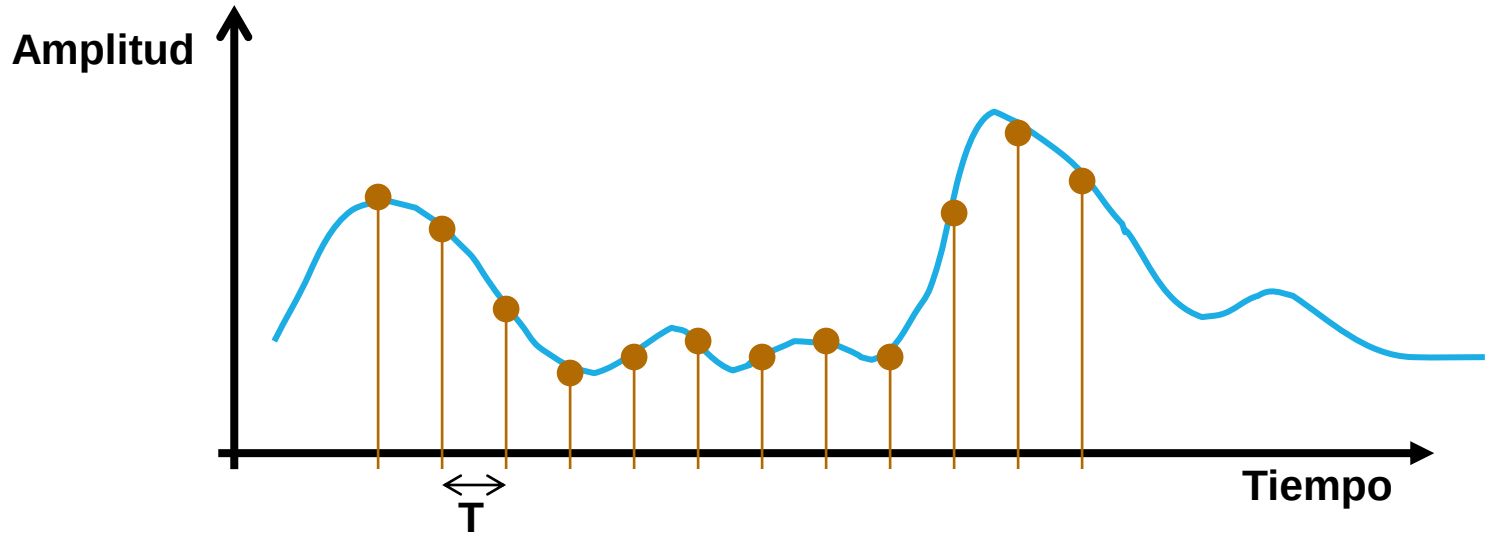
Audición humana



Tomado de:
"Psychoacoustics Facts and Models", Hugo Fastl and Eberhard Zwicker, Springer, 2007

Proceso de digitalización

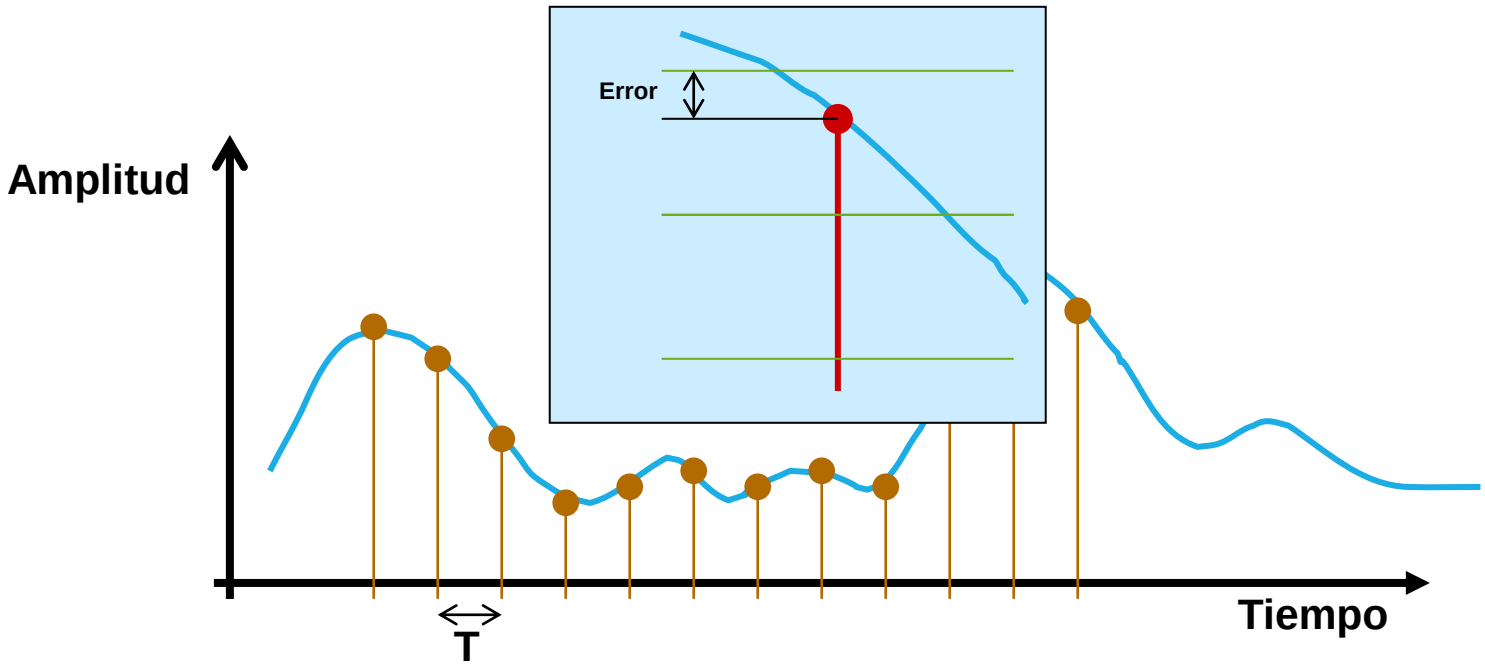
1. Muestreo
   - Se toman “muestras” de la señal a intervalos regulares. Estos intervalos deben ser tales que cumplan con el teorema de muestreo:
     - La mínima frecuencia a la que puede ser muestreada una señal y luego reconstruida es el doble de la frecuencia máxima de dicha señal



Proceso de digitalización

2. Cuantificación

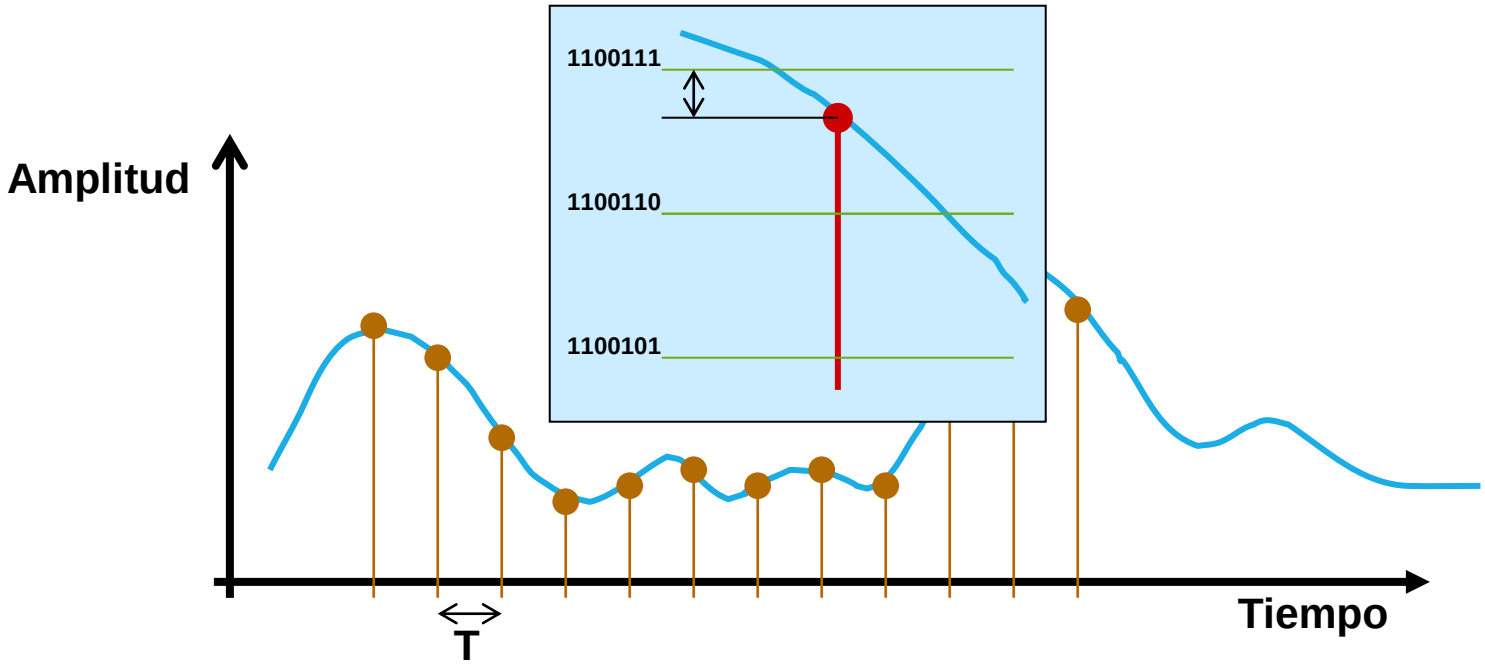
- Los valores de las muestras se “cuantifican” en cantidades discretas. La cantidad total de valores discretos debe introducir la menor cantidad posible de “ruido de cuantificación”



Proceso de digitalización

3. Codificación

- Los valores “cuantificados” se “codifican” en números que pueden ser luego procesados digitalmente.



Frecuencia de muestreo y bitrate

¿Cuántas muestras por segundo son necesarias?

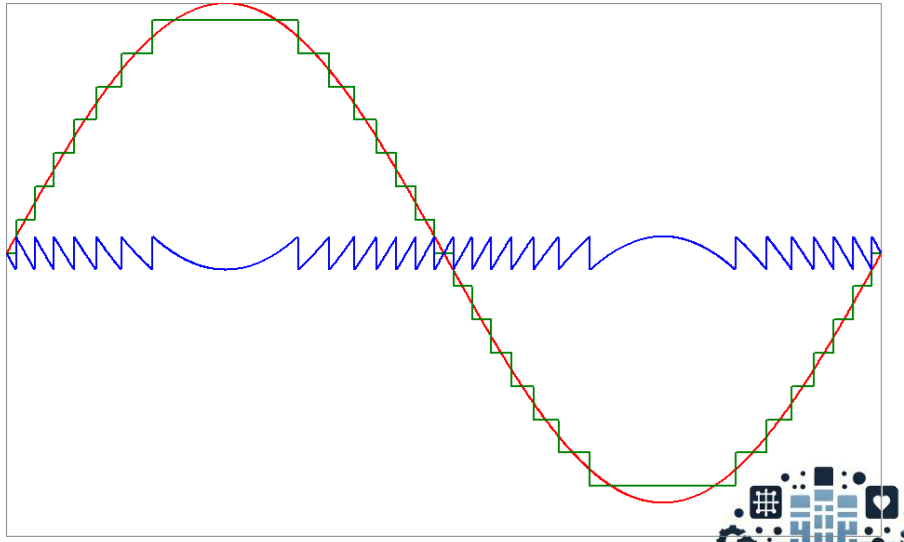
- Considerando que el oído puede escuchar hasta unos 20 kHz, y tomando en cuenta el *"teorema del muestreo"*, la frecuencia de muestreo debería ser > 40 kHz

¿Cuántos bits por muestra son necesarios?

- Al digitalizar, se introduce una distorsión, propia del proceso de cuantización

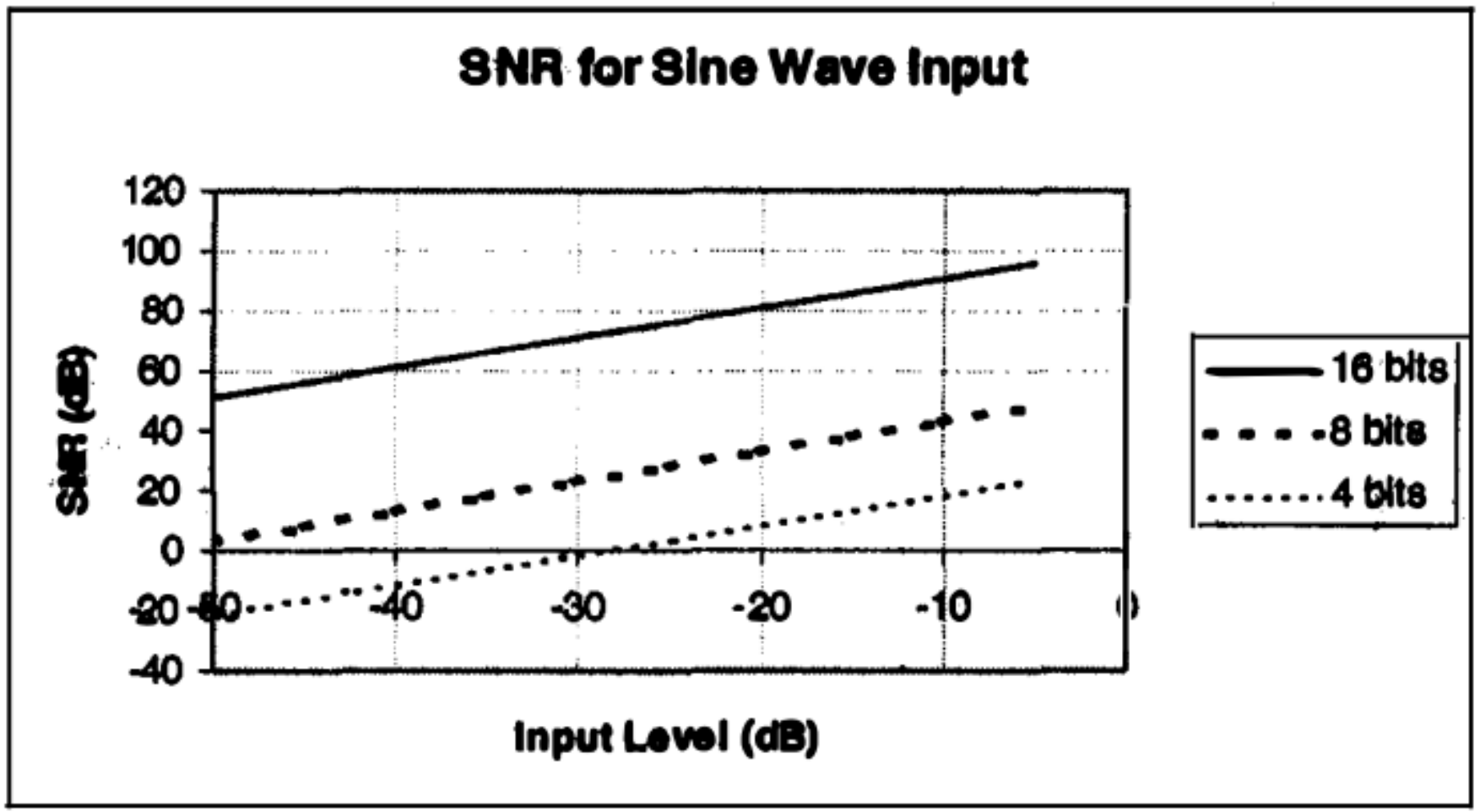
$$q(t) = x_{out}(t) - x_{in}(t)$$

$$SNR = 10log_{10}\left(\frac{< x_{in}^2 >}{< q^2 >}\right)$$



- Con más bits por muestra, menor distorsión

SNR según cantidad de bits, cuantización lineal



Tomado de:
"Introduction to Digital Audio Coding and Standards, Marina Bosi and Richard Goldberg, KLUWER ACADEMIC PUBLISHERS

Ejemplo: CD (Linear PCM)

44.1 kHz x 16 bits/muestra x 2 canales (stereo) =
1.4 Mb/s

Una canción de 10 minutos:

10 x 60 x 1.4 = 840 Mbits = 105 MBytes

Un CD tiene 700 MB, almacenaba unas 7 canciones

¿Cómo bajar el bitrate, y mantener una calidad aceptable?

¿Qué es aceptable?

- Depende de la aplicación…
  - Telefonía
  - Video conferencias
  - Conciertos
  - Música
  - …
  - ¿Mono, estéreo o más canales?

¿Hay otros factores?

- Retardos de codificación y de de-codificación
- Capacidad de procesamiento
- Robustez a los errores o pérdidas de información

Tipos de codecs

Pueden ser caracterizados por

- su tecnología
- su tasa de bits (bit rates)
- la calidad resultante del audio codificado
- su complejidad
- el retardo que introducen
- Otros factores…

Tipos de codecs

De voz

- Hacen uso de las características específicas de la voz humana
- Utilizados típicamente en aplicaciones de telefonía y de video conferencias
- Pueden ser de “forma de onda” o de “síntesis de voz”

De audio

- Hacen uso de las características más generales del aparato auditivo y la percepción humana del audio

Universales

Tipos de codecs

Codificación “con pérdida”

- Para comprimir “descartan” información que sea perceptualmente irrelevante o reiterativa
  - MP3 (MPEG-1 Audio Layer III), AAC (Advanced Audio Codec), Dolby AC-3, …

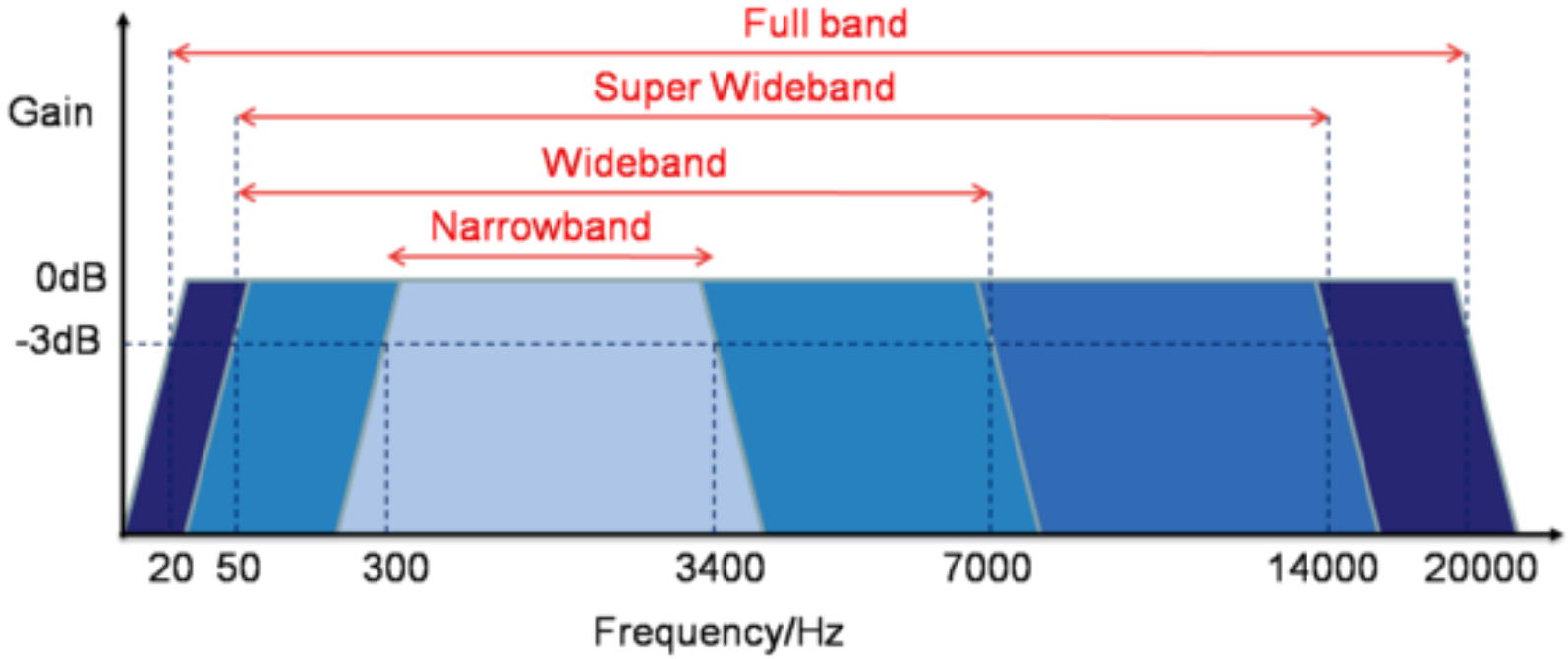
Codificación “sin pérdida”

- Comprimen sin perder información, se puede reconstruir exactamente el flujo original de bits
  - PCM, FLAC (Free Lossless Audio Codec), MPEG-4 ALS (Audio Lossless Coding”), ALAC (Apple Lossless Audio Coding), …

Tipos de codecs

Según el ancho de banda de la señal de entrada

- Banda angosta (narrowband) 300 a 3400 Hz
- Banda ancha (wideband) 50 a 7000 Hz
- Banda super ancha (superwideband) 50 a 14000 Hz
- Banda completa (fullband) 20 a 20000 Hz



Codificación de Voz

CODECs de banda angosta

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
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
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)
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
AMR	Adaptive Multi Rate	12.2 a 4.75	20	Utilizado en redes celulares GSM

CODECs de banda ancha

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 de telefonía de calidad en VoIP
G.722.1	Transform Coder	24,32	40	Usado en audio y videoconferencias
G.722.2	AMR-WB	6.6 a 23.85	25.9375	Estandar en común con 3GPP (3GPP TS 26.171). gran inmunidad a los ruidos de fondo en ambientes adversos (por ejemplo celulares)
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
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
RtAudio	Real Time Audio	8.8, 18	40	Codec propietario de Microsoft, utilizado en aplicaciones de comunicaciones unificadas (OCS)

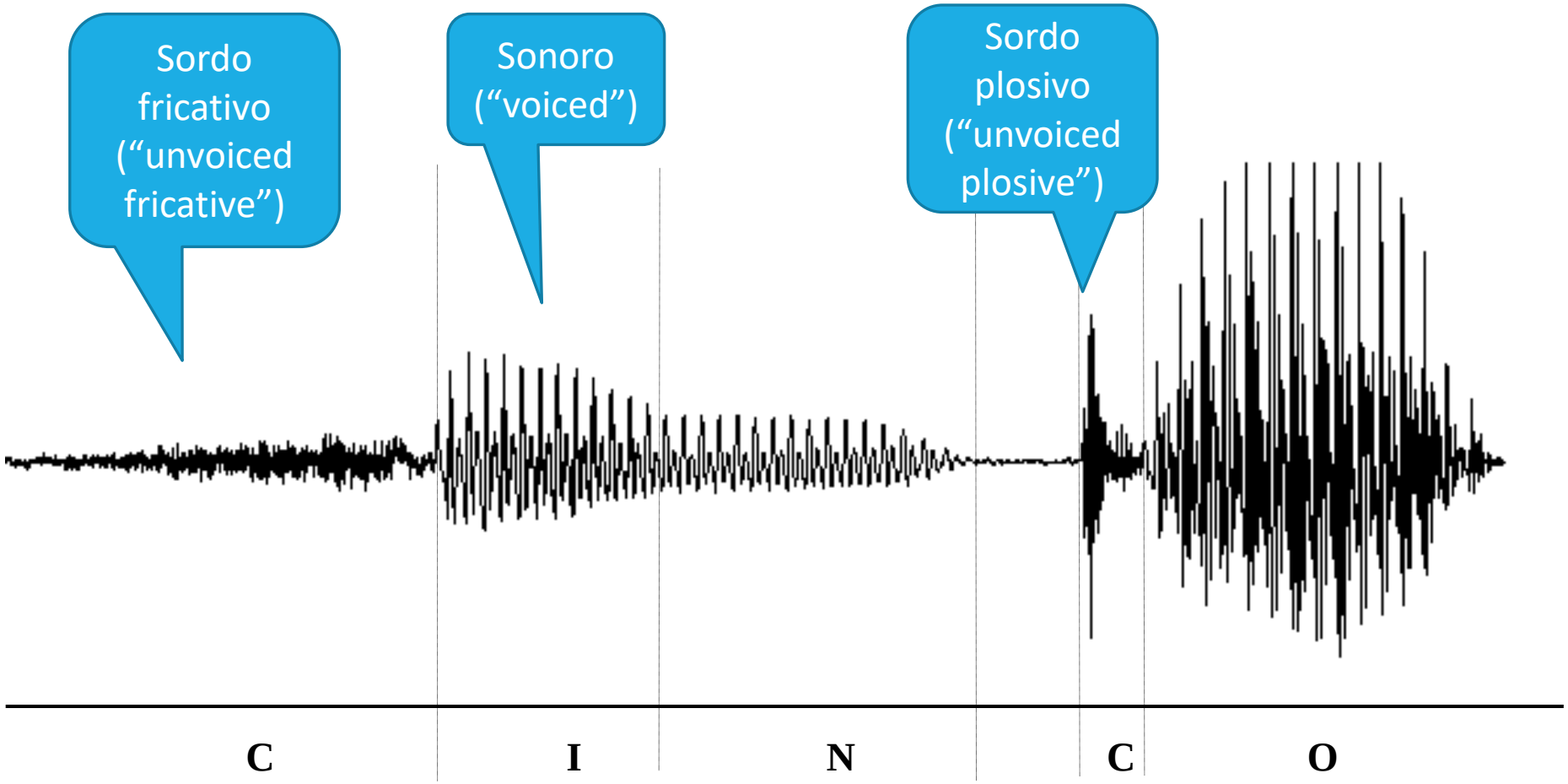
CODECs de banda superancha

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

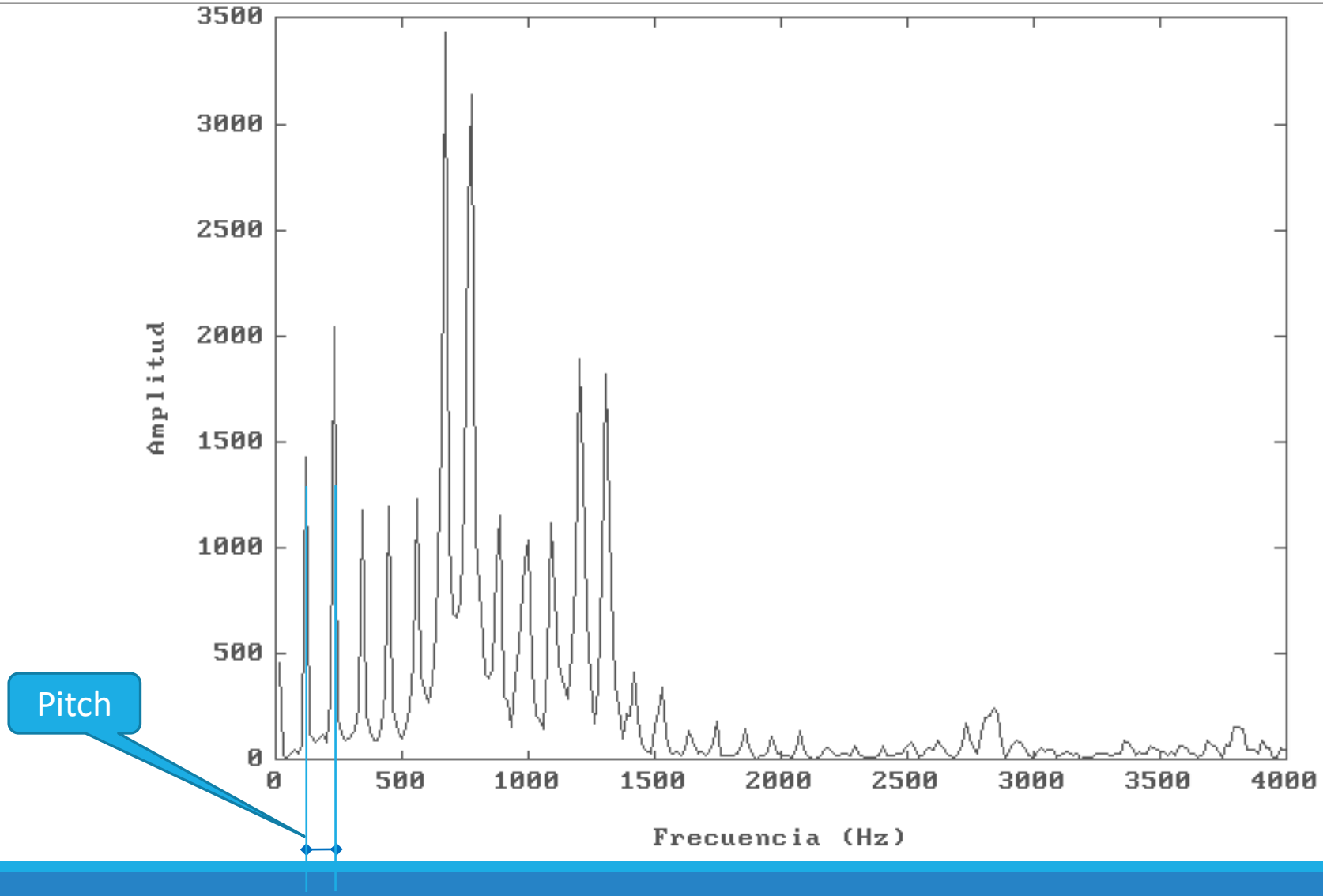
CODECs de banda completa

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
Opus	Opus	6 a 510	Hasta 60	Incorpora tecnología de SKYPE RFC 6716 (propuesta en set 2012)
Satin	Satin	6 a 36 (48 en el futuro)		Utilizado por Microsoft en su producto Teams

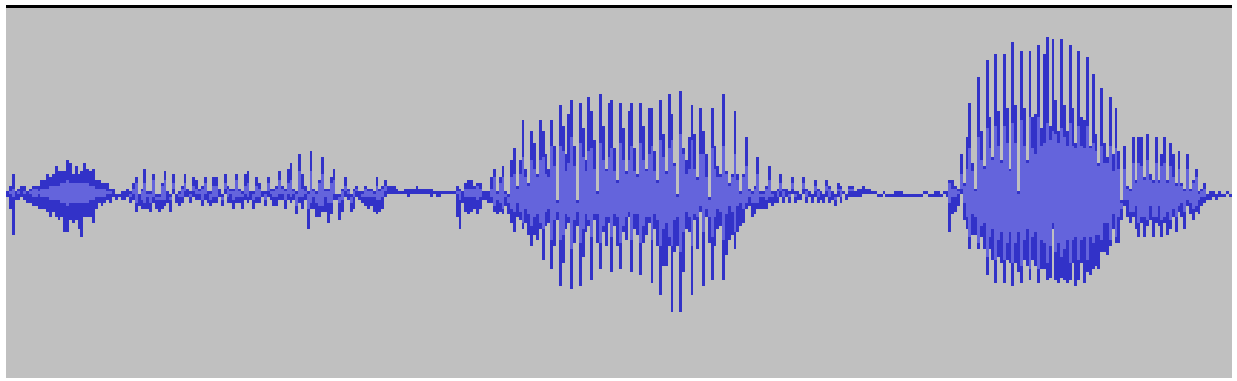
Características de la señal de voz

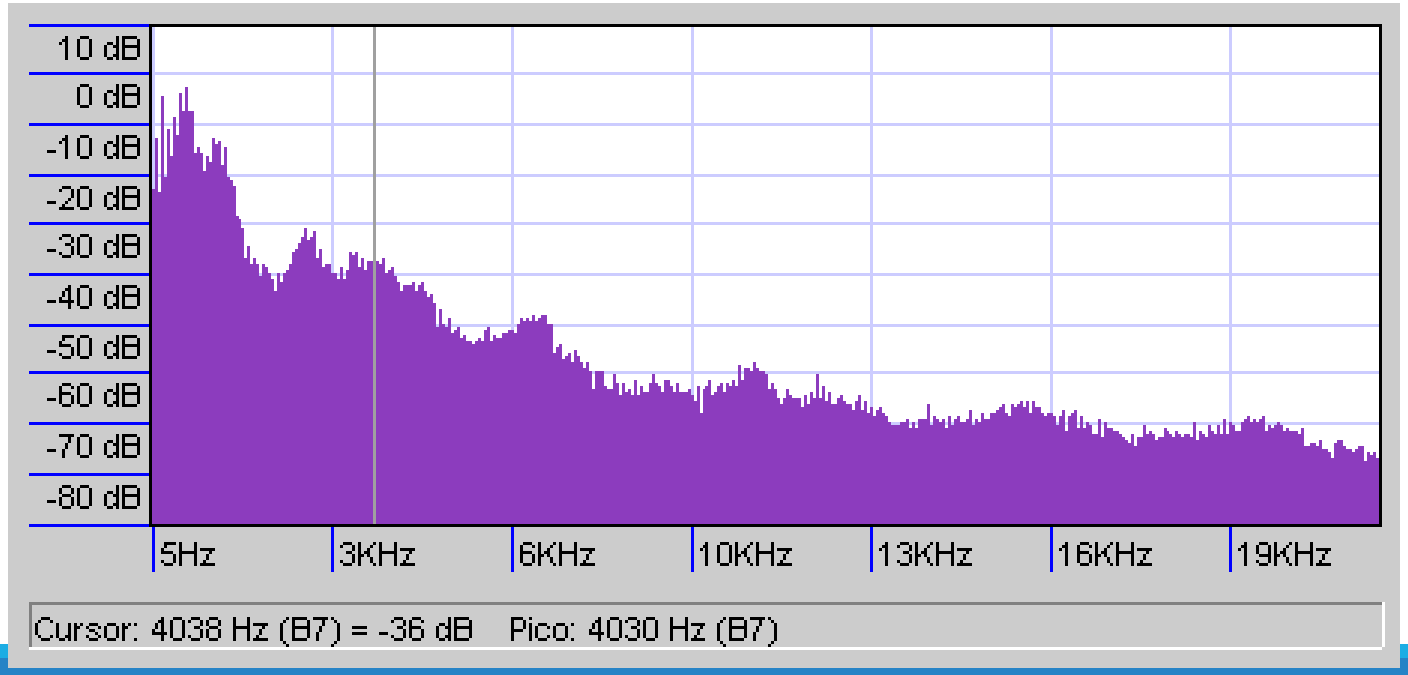


Espectro de la “A”



Espectro típico de la voz





Codificación de “forma de onda”

Inicialmente, los codecs se basaron en codificar de la manera más eficiente posible la “forma de onda” de la señal.

Posteriormente, para bajar la tasa de bits necesaria para la transmisión, se comenzaron a utilizar técnicas “predictivas”

- Basadas en predecir los valores de las muestras en base a la extrapolación de las muestras anteriores

G.711 – Pulse Code Modulation (PCM)

Primer codec de Voz, diseñado para telefonía

Estandarización de “Ley A” y “Ley Mu”

Conserva la forma de onda, codifica muestra a muestra

Tiene características “no lineales” para minimizar la cantidad de bits por muestra

Resulta en una velocidad de **64 kbit/s**

G.711

1. Muestreo

- Si bien el oido humano puede llegar a escuchar sonidos de hasta 18 - 20 kHz, la mayor parte de la energía de la voz humana se encuentra por debajo de los 4 kHz.
- El sonido resultante de filtrar la voz humana a 3.4 kHz es perfectamente inteligible, además puede distinguirse al locutor.
- De acuerdo al teorema del muestreo, para poder reconstruir una señal de 3.4 kHz debe muestrarse a más de 6.8 kHz.
- Originialmente se seleccionó como frecuencia de muestreo para telefonía **8 kHz** (una muestra cada **125** microseg).

G.711

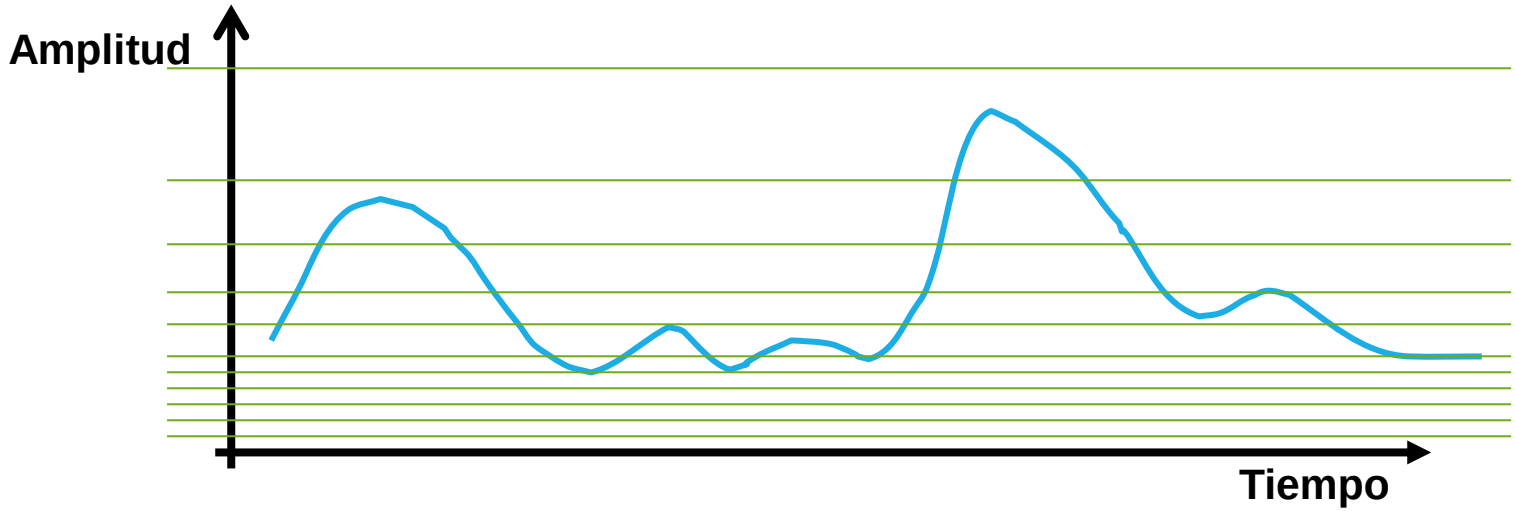
2. Cuantificación (1/3)

- Una cuantificación lineal genera un “error de cuantificación” constante, independiente del nivel de la señal.
- Los errores de cuantificación se traducen en “ruido” al reconstruir la señal.
- Para lograr niveles de ruido aceptables en señales de voz con cuantificadores lineales, se requieren 4096 niveles.
- El oído es mas sensible a los “ruidos” en señales bajas que en señales altas.

G.711

2. Cuantificación (2/3)

- Cuantificación no lineal: Permite tener errores de cuantificación pequeños para señales pequeñas y grandes para señales grandes
- Con menos cantidad de niveles se logra buena calidad en la señal reconstruida



G.711

2. Cuantificación (3/3): Leyes de Cuantificación

- Ley A (de 13 segmentos):

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

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

  $A = 87.6$

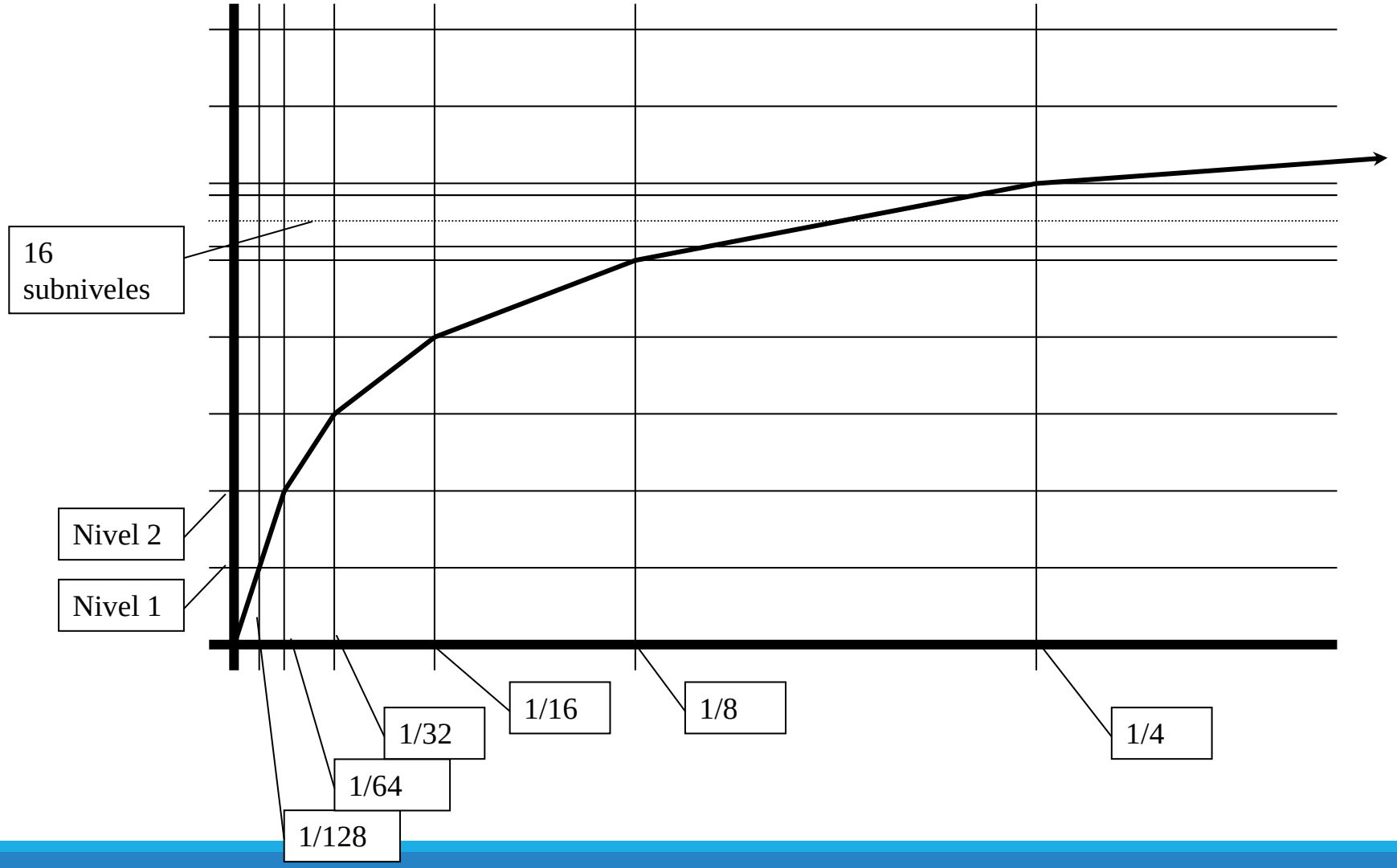
- Ley μ (de 15 segmentos):

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

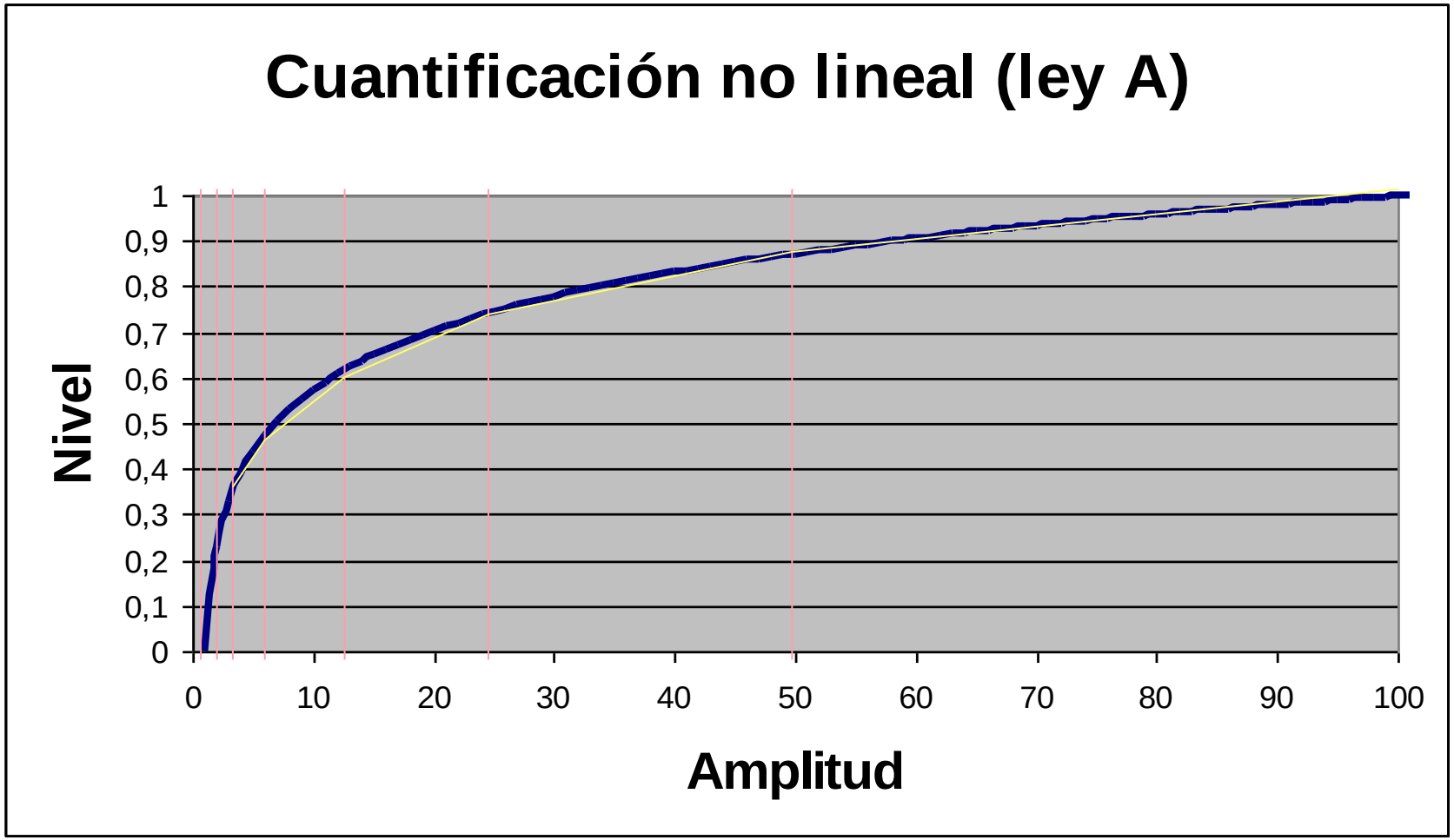
  $\mu = 255$

- Ambas leyes forman parte de la Recomendación ITU-T G.711

G.711 - Ley A



G.711 - Ley A


Cuantificación no lineal (ley A)
Nivel
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
10
20
30
40
50
60
70
80
90
100
Amplitud

G.711 - Ley A

3. Codificación: Ley A o ley de los 13 segmentos

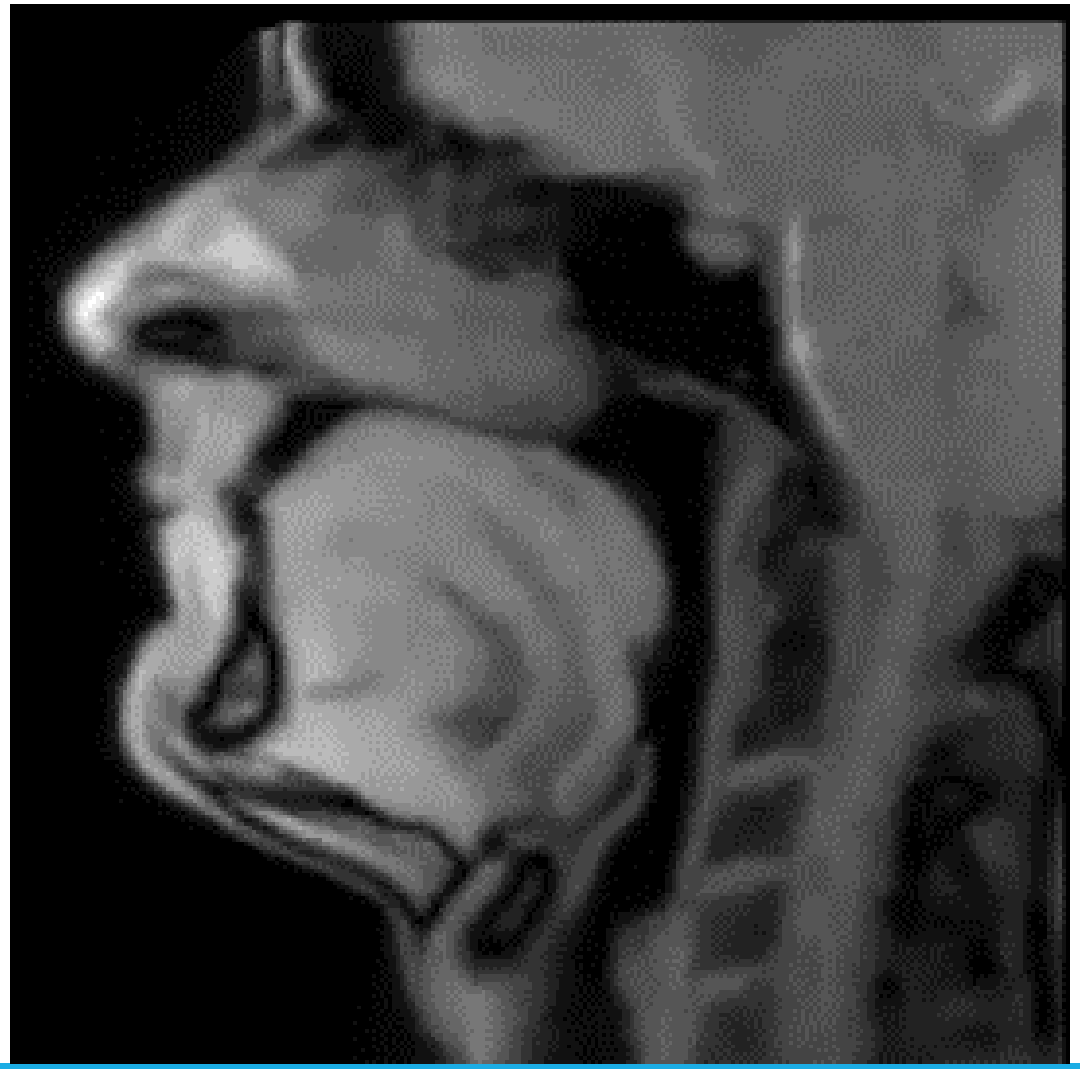
- El bit más significativo (bit 7) indica el signo.
- Los bits 4-6 indican el número de segmento.
- Los bits menos significativos (bits 0-3) indican el intervalo dentro del segmento.

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

Codificación por “síntesis de voz”

Video de cuerdas vocales

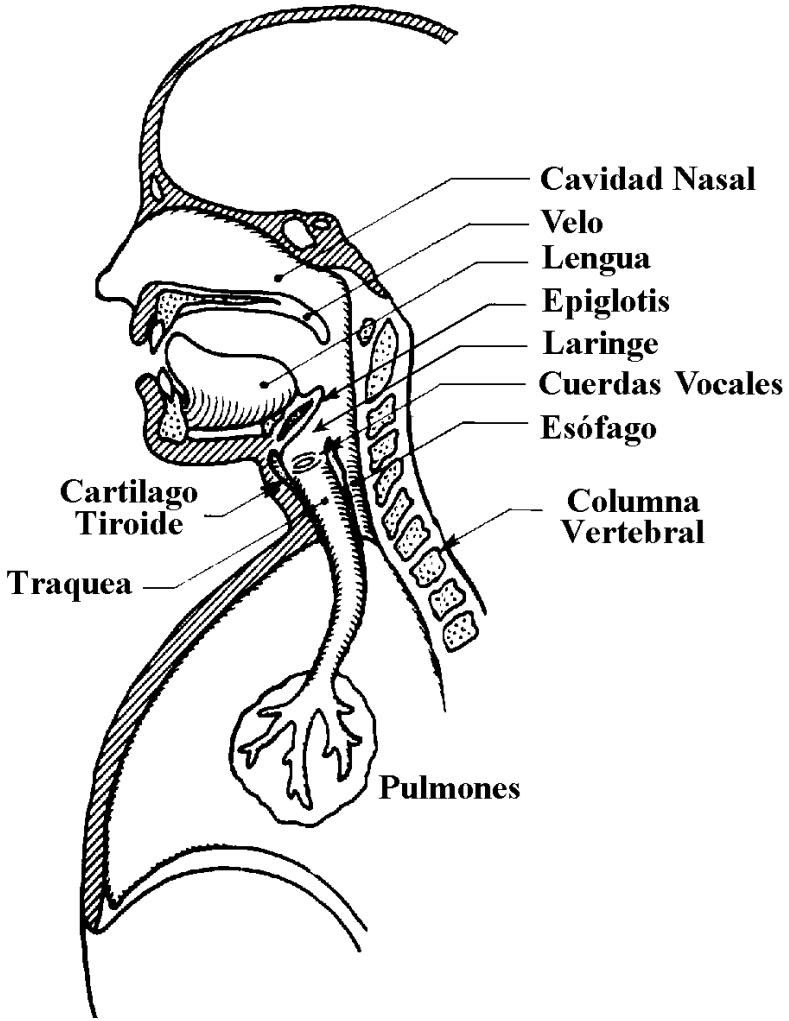
Video de cuerdas vocales



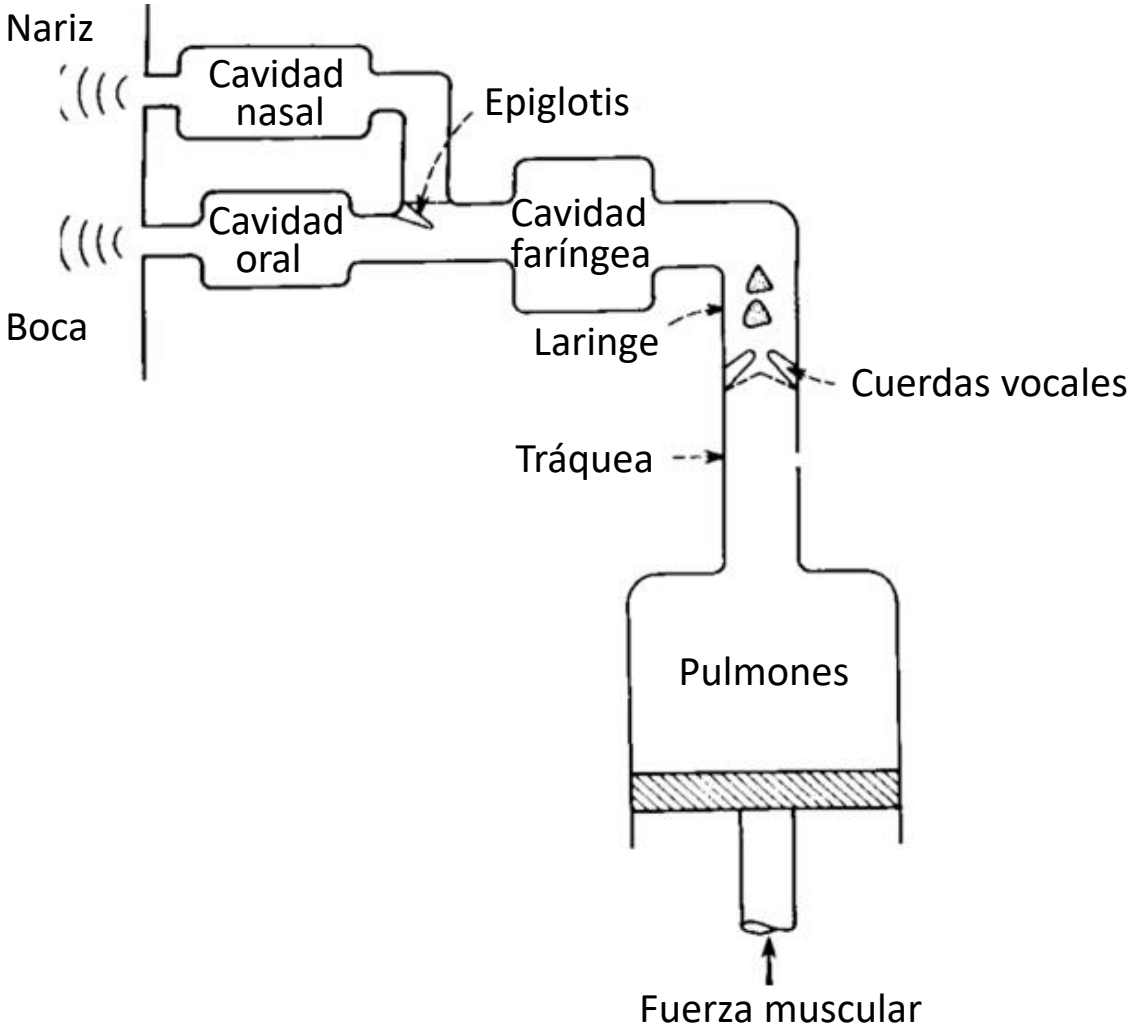
De. Jens Frahm / MPIBPC (https://www.mpinat.mpg.de/626786/real-time-mri)

Codificación por “síntesis de voz”

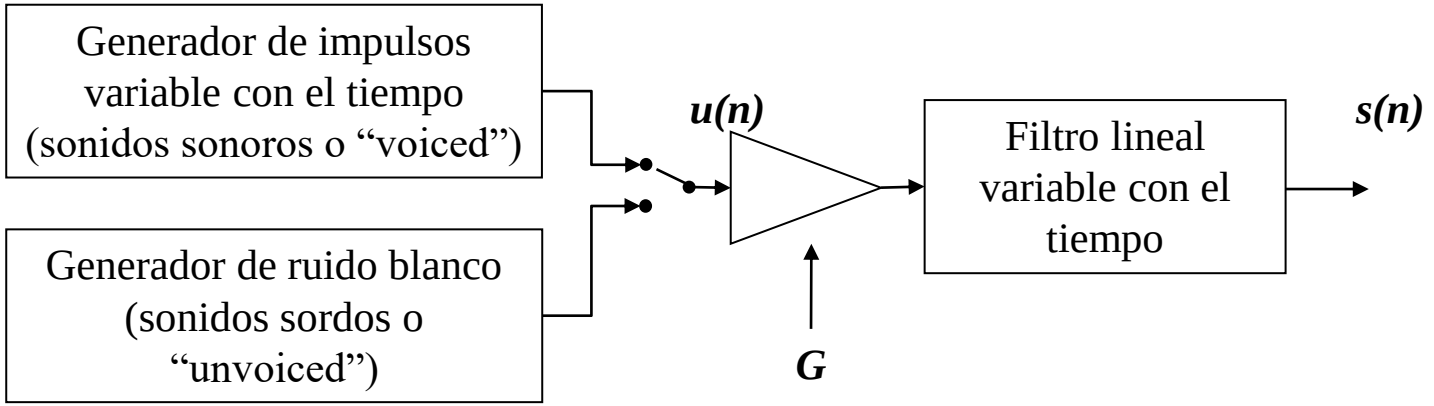
Sobre la década de 1980, se introduce la idea de generar “voz sintética”, simulando la manera en que se produce la voz humana en el conducto vocal.



Modelo del Conducto Vocal



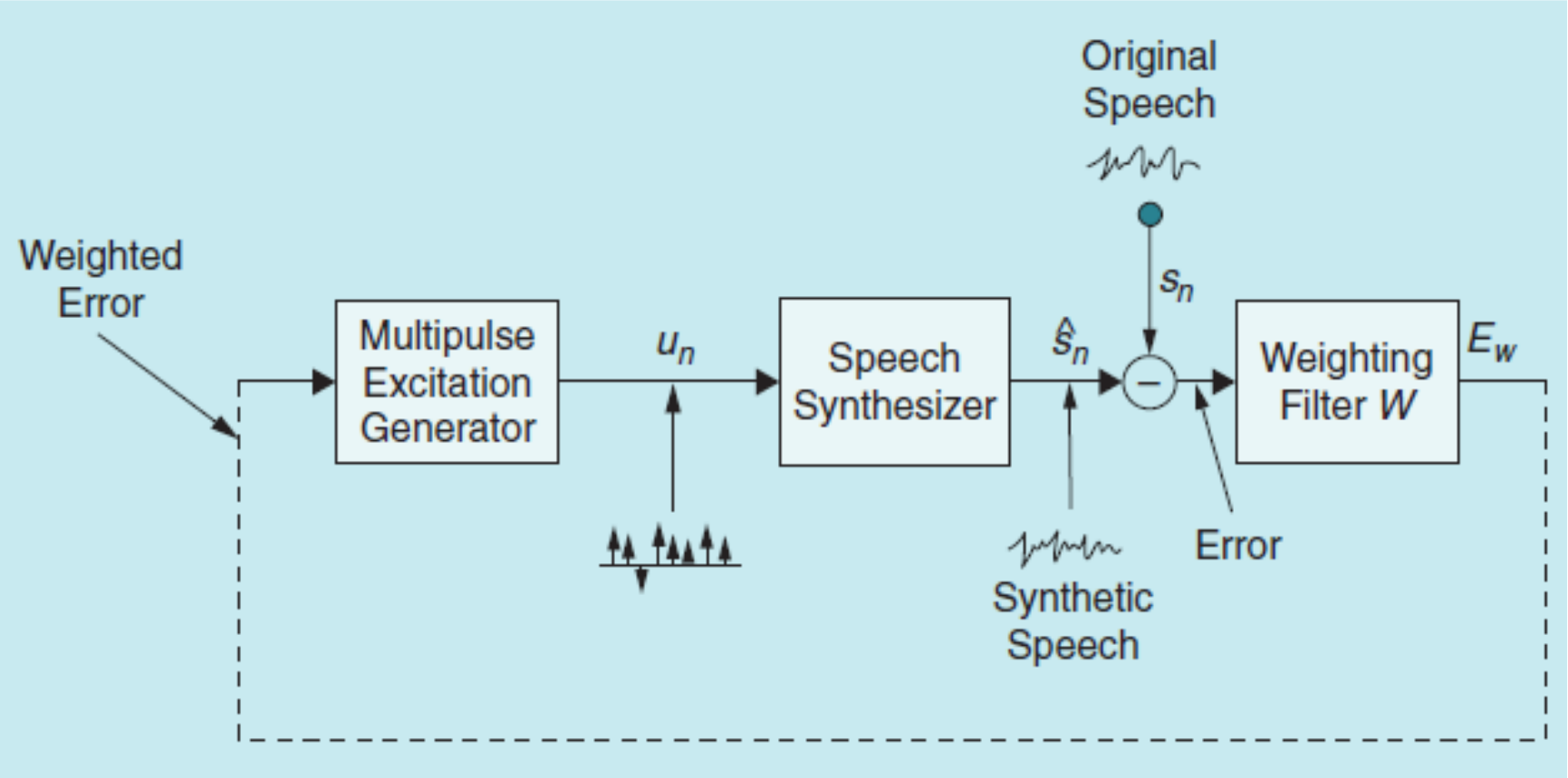
Modelo del Conducto Vocal



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

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

Estimación de los parámetros del modelo



Tomado de:
The History of Linear Prediction, Bishnu S Atal, IEEE Signal Processing Magazine, March 2006, pp 154-161

G.729

Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear prediction (CS-ACELP)

No conserva la forma de onda, sino que utiliza técnicas de “síntesis de voz”

El modelado de la boca y la garganta se hace por medio de filtros lineales y la voz se genera a partir de una vibración periódica de aire que los excita

Utiliza “ventanas” de 10 ms para obtener los parámetros y se usan 80 bits (10 bytes) para representarlos

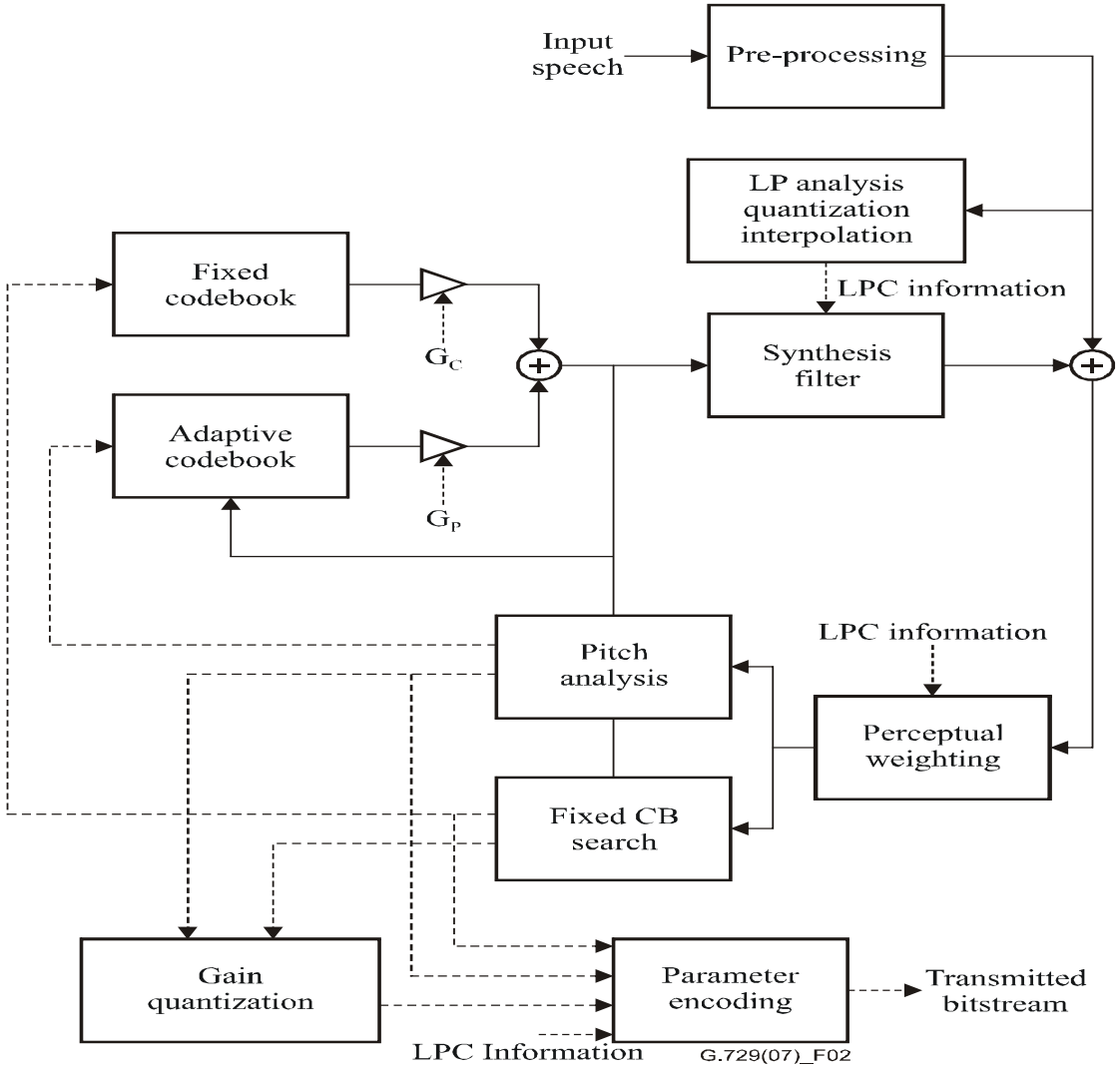
- Resulta en una velocidad de 8 kbit/s

G.729

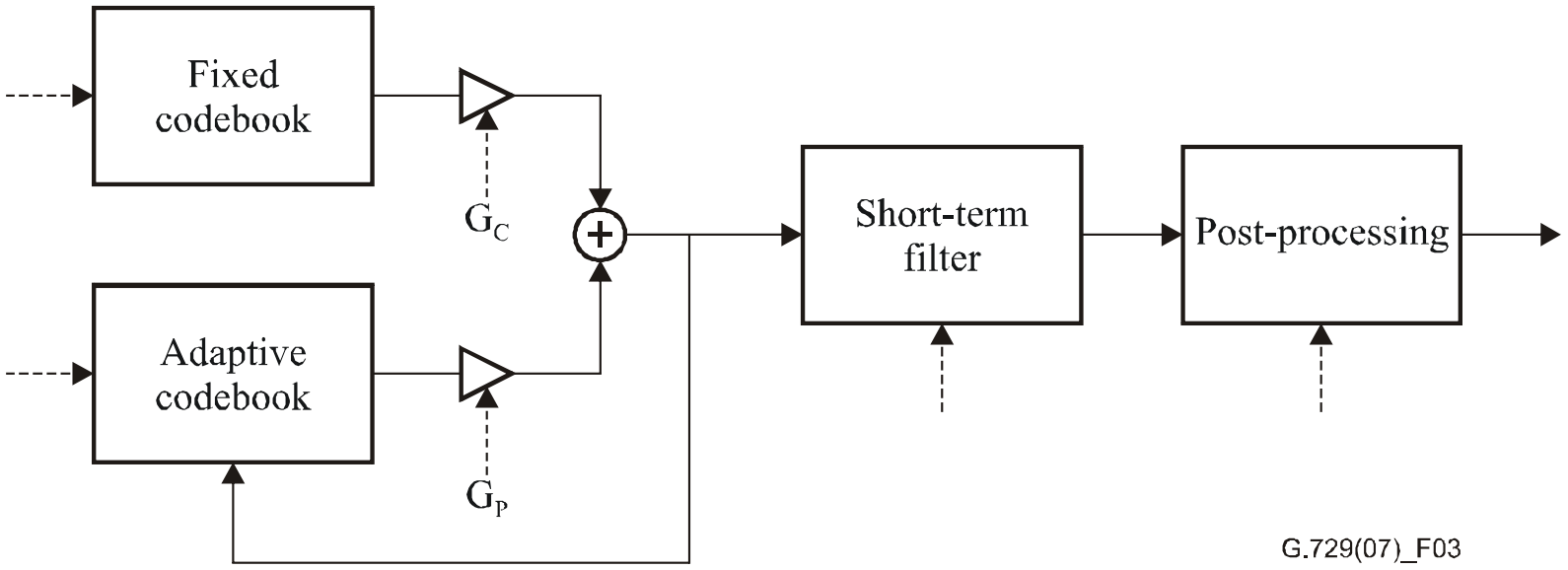
Tiene 5 ms de “look ahead”, resultando en una demora total de 15 ms

Utiliza técnicas CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear Prediction)

Codificador G.729



Decodificador G.729



G.729

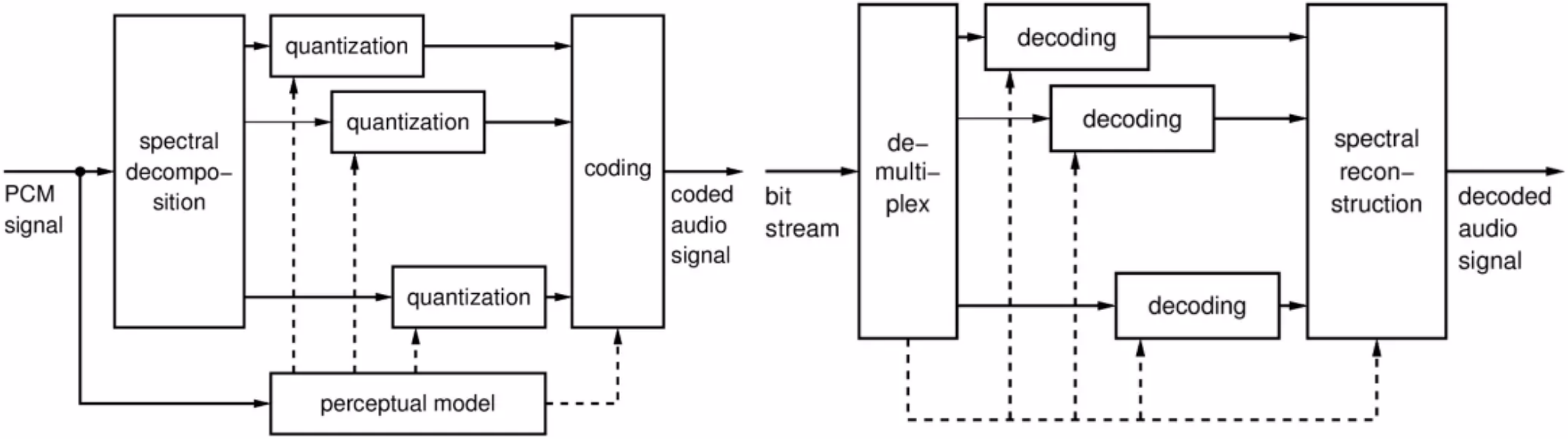
G.729 A

- Variante del codec para lograr menor complejidad
- Es interoperable con G.729

G.729 B

- Detección de actividad de voz y silencios
- Modelado y regeneración del “ruido de fondo” (CNG = Confort Noise Generation)
- Menor ancho de banda en la LAN

Codificación por descomposición espectral



Encoder

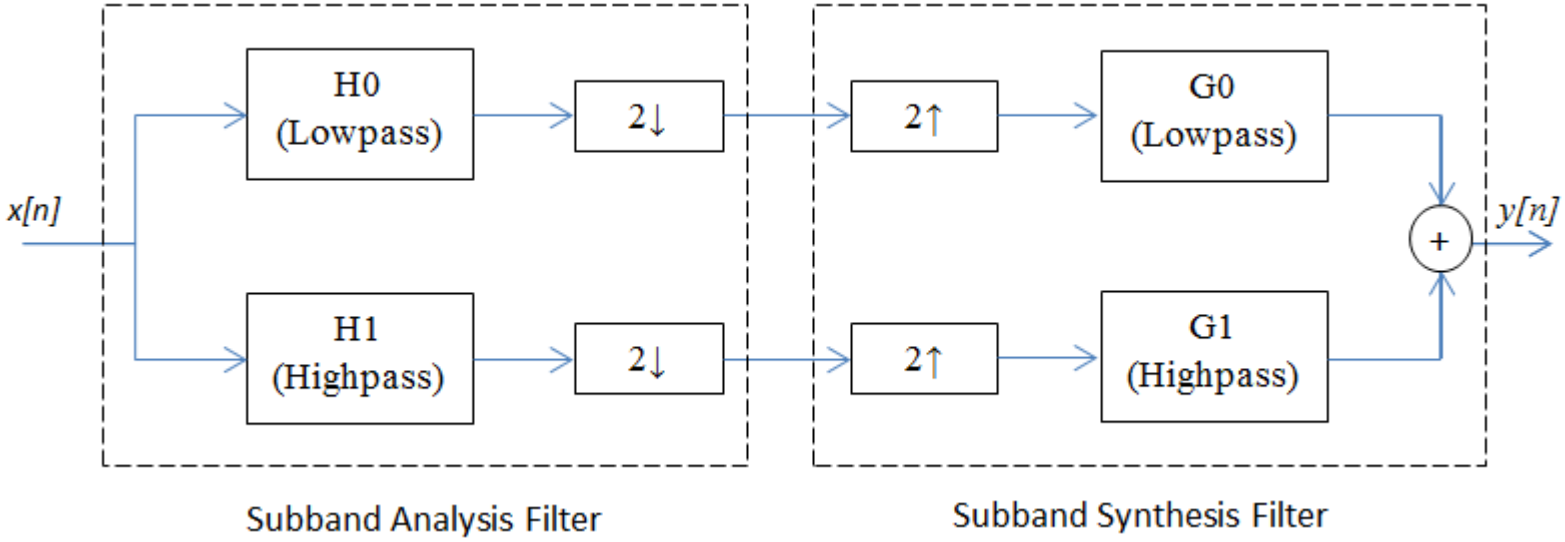
Decoder

Tomado de:
The MDCT and its applications in audio coding, IMTC anual member meeting 2014, Bernd Edler

Codificación por descomposición espectral

Quadrature Mirror Filter (QMF)

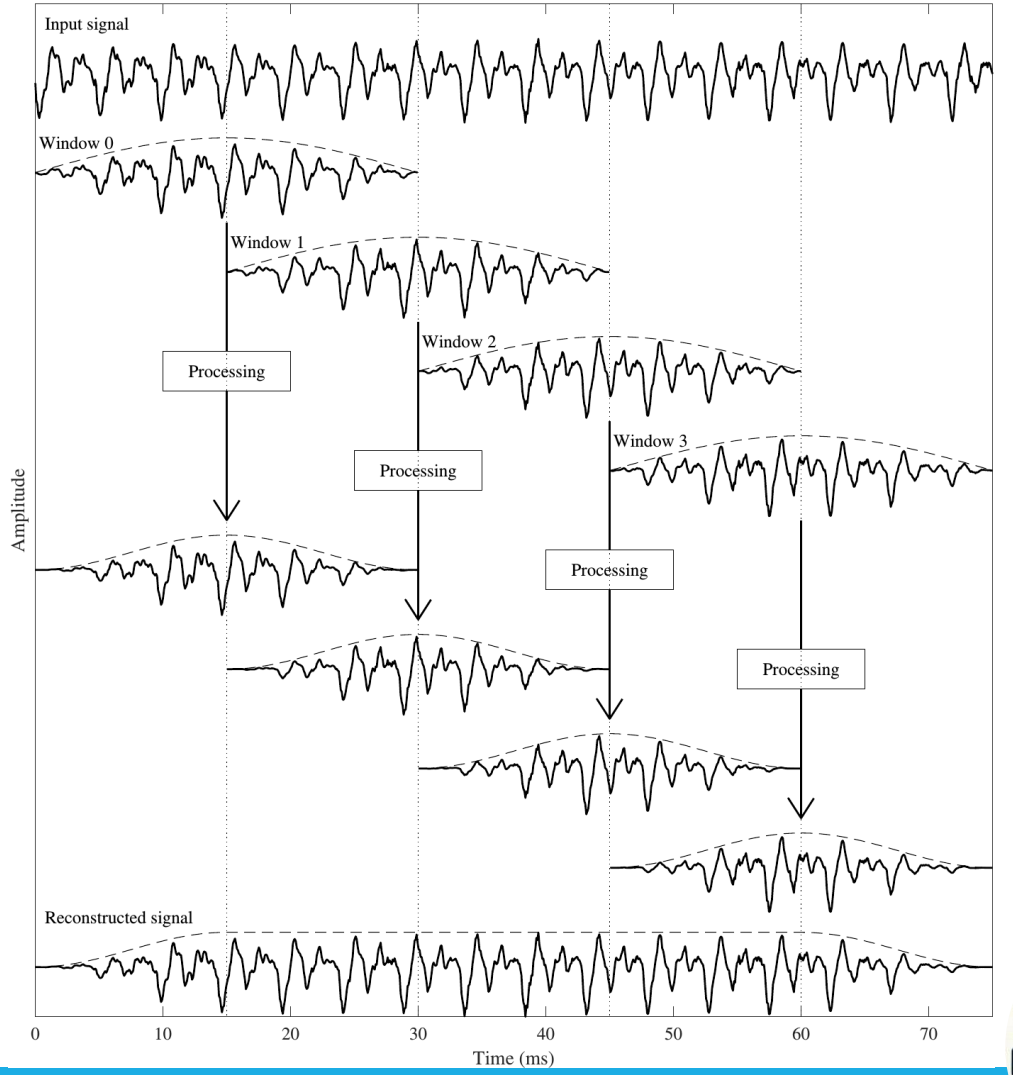
El banco de filtros QMF consta de filtros de análisis que descomponen la señal en dos sub-bandas



Codificación por descomposición espectral

Modified Discrete Cosine Transform (MDCT)

Separa la señal en bloques superpuestos. A diferencia de la DCT, la superposición de las ventanas evita distorsiones, pero a su vez, no implica la generación de más bits de codificación.



De:
https://speechprocessingbook.aalto.fi/Transmission/Modified_discrete_cosine_transform_MDCT.html

G.711.1 - Wideband embedded extension for G.711 PCM

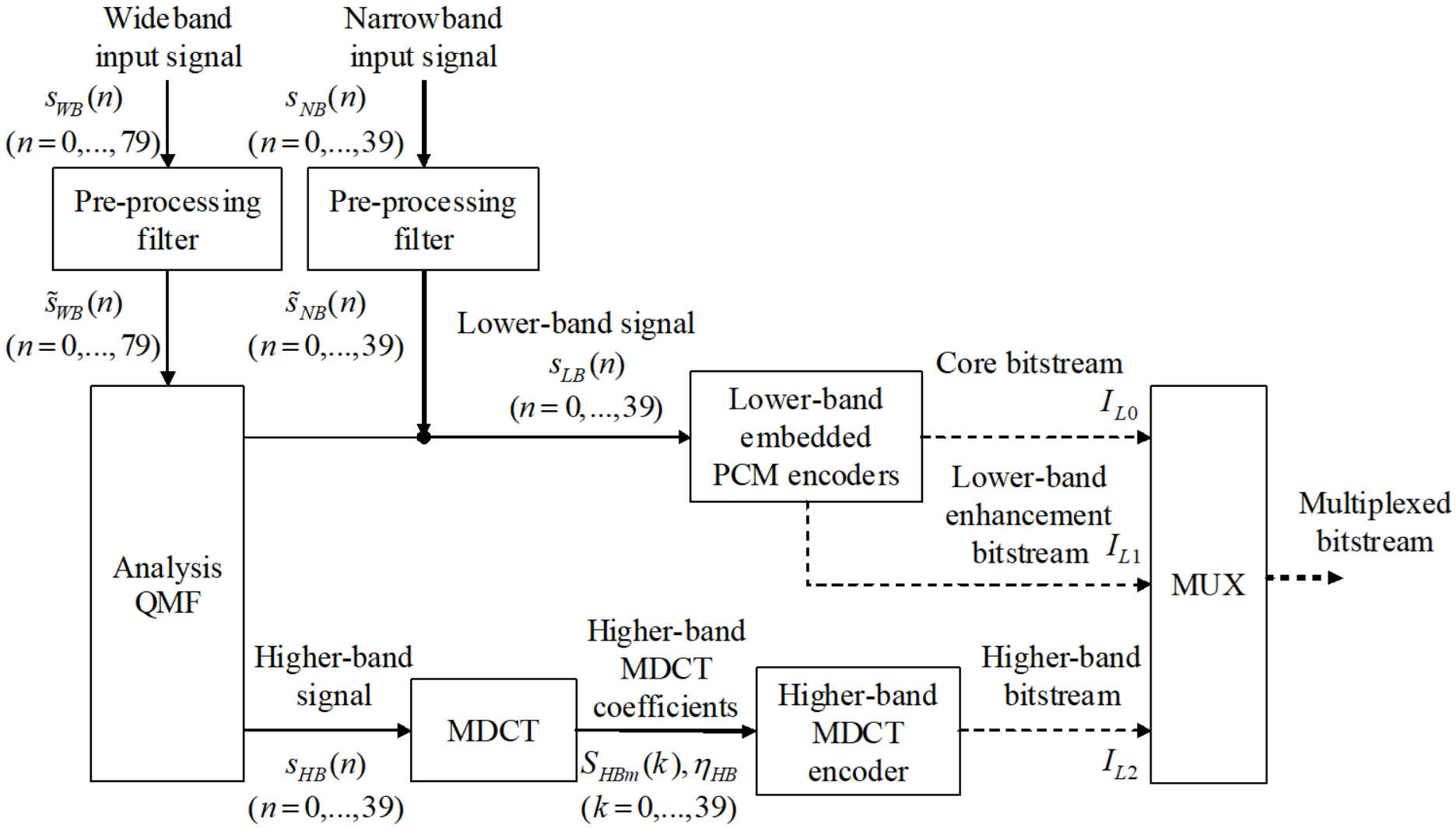
Aprobado en Marzo de 2008, como una extensión de G.711 para banda ancha (7 kHz)

Trabaja en 64, 80 y 96 kb/s

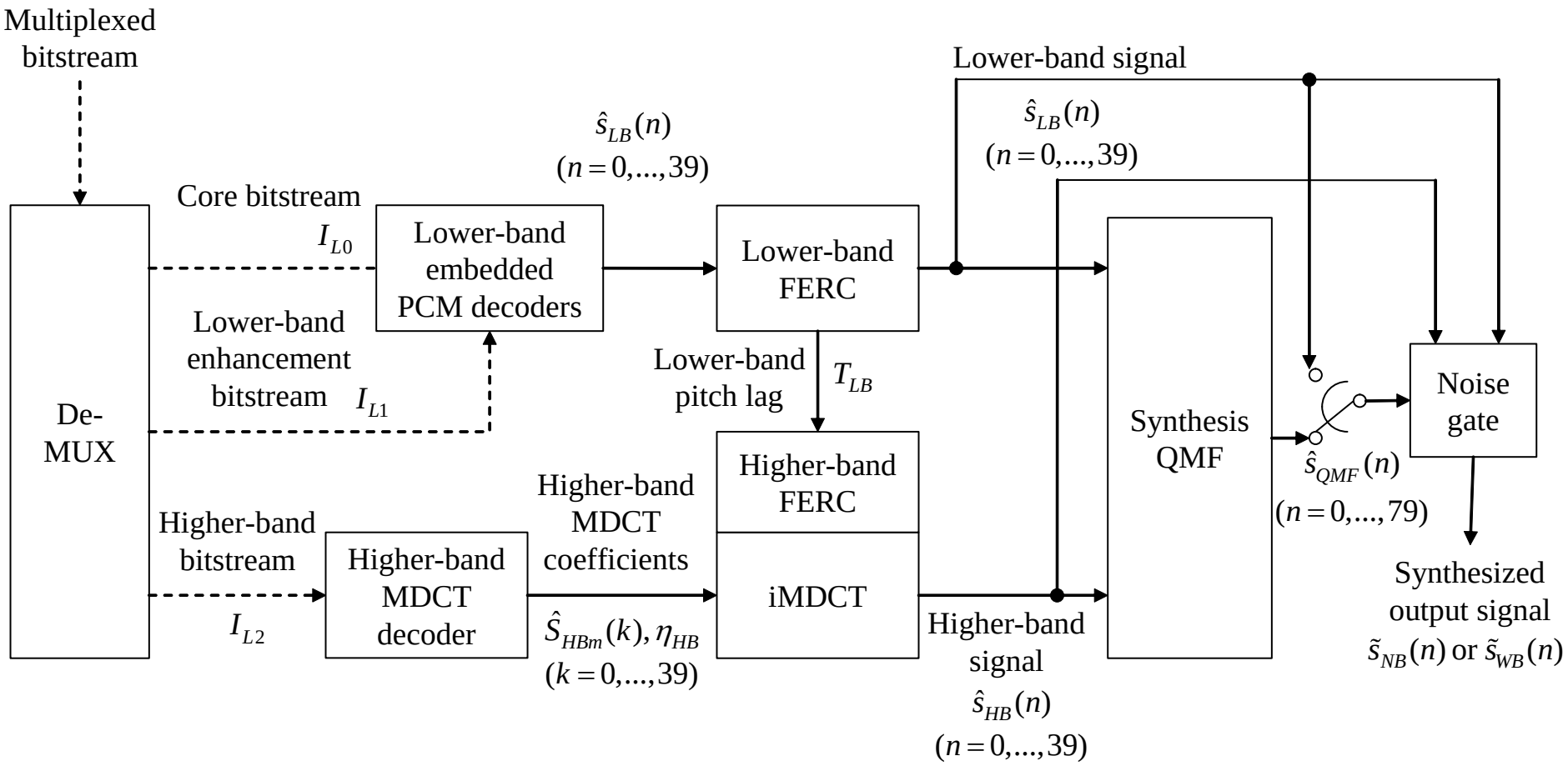
Las muestras codificadas pueden ser convertidas en G.711 por medio de un simple truncado

Las muestras de entrada son tomadas cada 16 kHz, pero también está soportada la frecuencia de muestreo de 8 kHz (compatibilidad con G.711)

Codificador G.711.1



Decodificador G.711.1



Modos de operación G.711.1

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

Tramas G.711.1

Son de 5 ms y tienen un total de 480 bits por trama

- 320 bits de la capa 0 (G.711), correspondientes a 8 bits x 40 muestras
- 80 bits de la capa 1
- 80 bits de la capa 2

La demora total del algoritmo lleva un total de 11.875 ms

- 5 ms para la información de la trama
- 5 ms extras necesarios para el análisis MCDT (“lookahead”)
- 1.875 ms para la implementación del filtro QMF

Comfort Noise Generation

Aproximadamente, durante el 40% de una conversación multimedia, “escuchamos sin hablar”

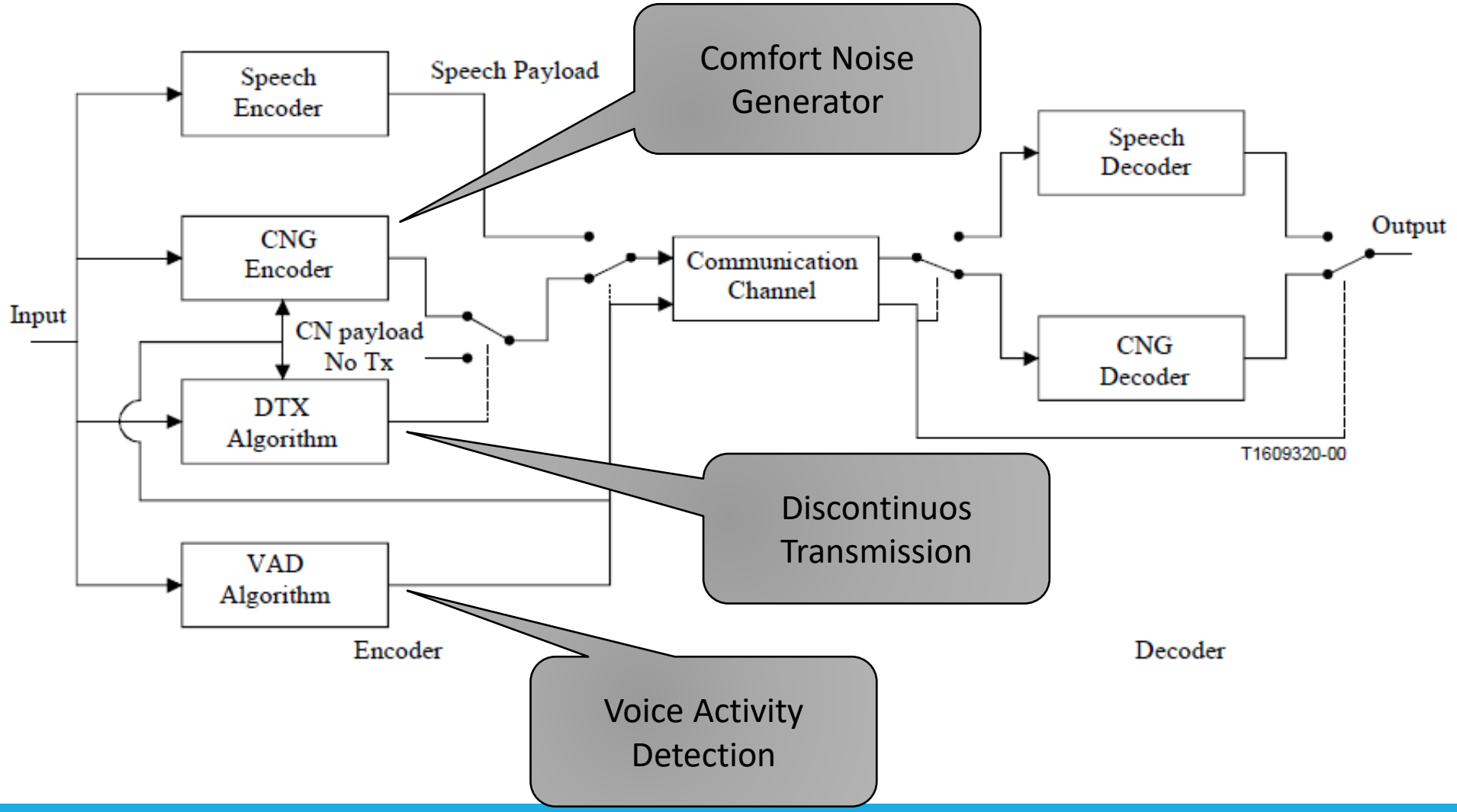
La tecnología de Comfort Noise Generation (CNG) se utiliza en sistemas de comunicación para generar un “ruido de confort sintético” que se introduce cuando no hay señal de voz activa.

Este ruido artificial se implementa para evitar que los periodos de silencio sean percibidos como desconexiones o interferencias en la comunicación.

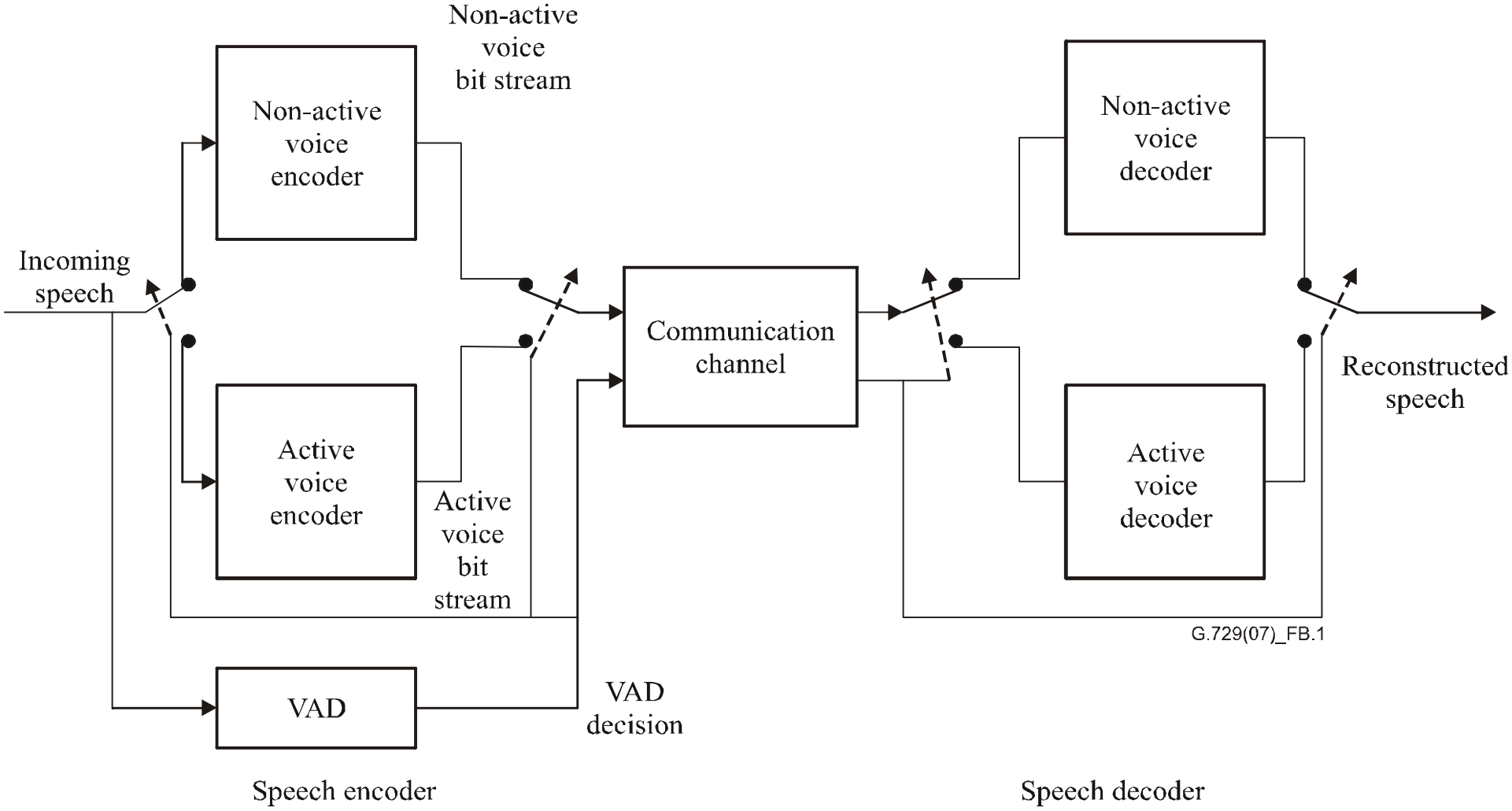
La generación de este ruido se basa en reproducir el ambiente sonoro durante los momentos de inactividad vocal, lo que contribuye a mantener una experiencia continua y natural para los usuarios.

Además, el CNG se adapta al entorno acústico y a la calidad de funcionamiento del sistema, promediando la estimación del ruido ambiental y ajustando su carácter espectral al del códec vocal utilizado

G.711 Appendix II – Comfort Noise Generation



G.729 B VAD (Voice Activity Detection)



G.729.1

An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729

Aprobado en mayo de 2006

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)

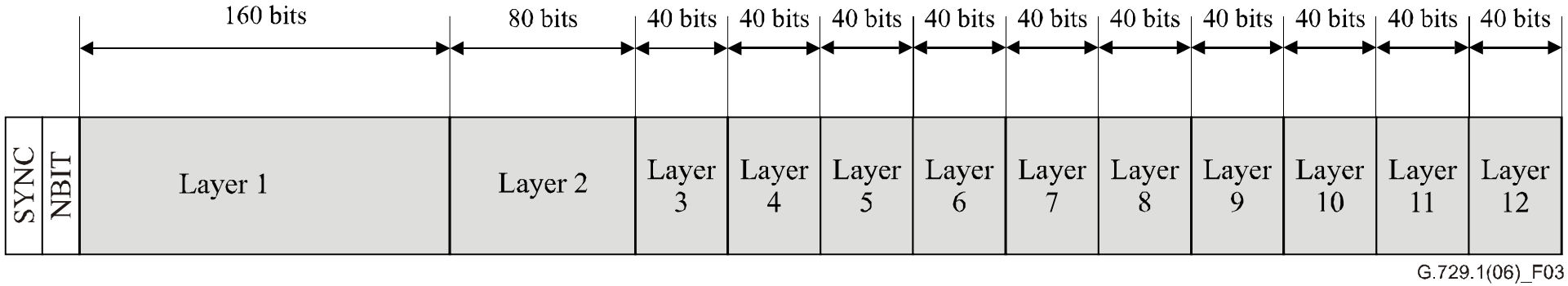
Inter operable con la recomendación G.729 y sus anexos A y B, los que tienen amplia difusión en el mundo de VoIP

Trama G.729.1

Capa 1: Codificación basada en CELP, de 8kb/s y compatible con G.729

Capa 2: Mejoras en las frecuencias de la banda baja (50 a 4000 Hz), de 4 kb/s

Capas siguientes: Agregan progresivas mejoras en la banda alta, 2 kb/s adicionales cada una



G.723.1

6.4 kb/s

- Utiliza un algoritmo MPC-MLQ (Multi-Pulse Maximum Likelihood Quantization), generando 24 bytes por cada ventana de 30 ms.

5.3 kb/s

- Utiliza ACELP (Algebraic Code Excited Linear Prediction), generando 20 bytes por cada ventana de 30 ms

El retardo total (latencia) es de 37.5 ms

- El algoritmo requiere de 7.5 msegundos de muestras adicionales (“look ahead”).

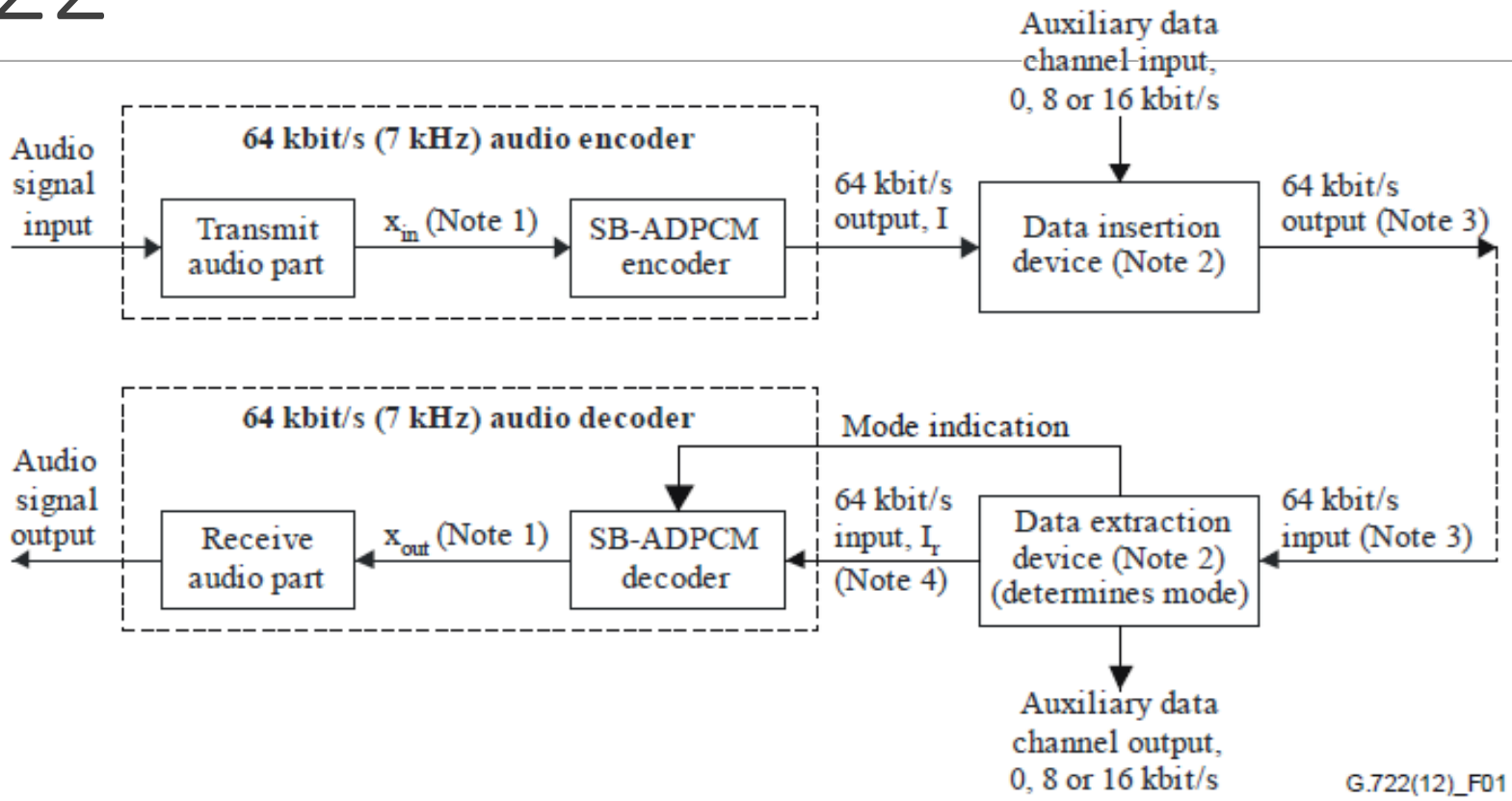
G.722 - 7 kHz audio-coding within 64 kbit/s

Codec de Banda Ancha

Utiliza técnicas de ADPCM, separando la señal en dos sub-componentes (banda baja y banda alta)

Opera en tres posibles modos, en 64, 56 o 48 kb/s

G.722



NOTE 1 – X_{in} and X_{out} are digital signals uniformly coded with 14 bits and 16 kHz sampling.
NOTE 2 – These devices are only necessary for applications requiring an auxiliary data channel within the 64 kbit/s.
NOTE 3 – Comprises 64, 56 or 48 kbit/s for audio coding and 0, 8 or 16 kbit/s for data.
NOTE 4 – 64 kbit/s signal comprising 64, 56 or 48 kbit/s for audio coding depending on the mode of operation.

Figure 1 – Simplified functional block diagram

G.722

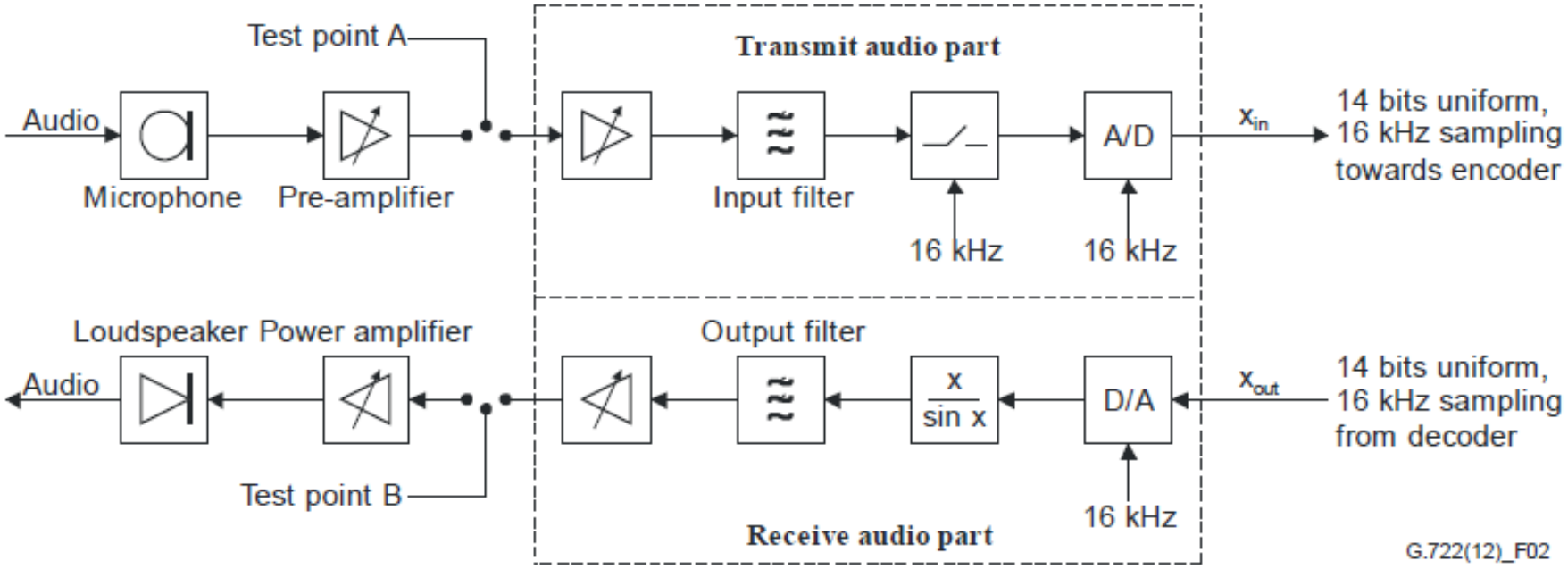


Figure 2 – Possible implementation of the audio parts

G.722

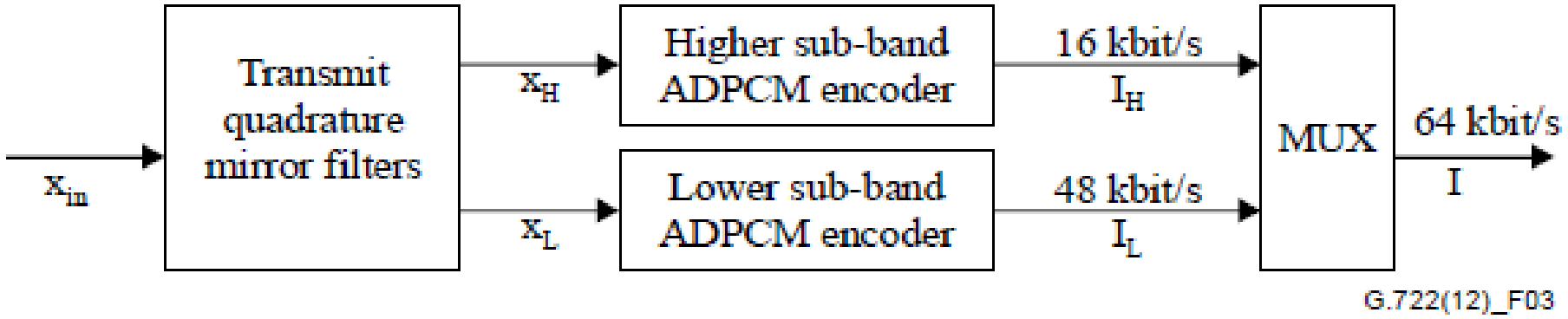


Figure 3 – Block diagram of the SB-ADPCM encoder

G.722

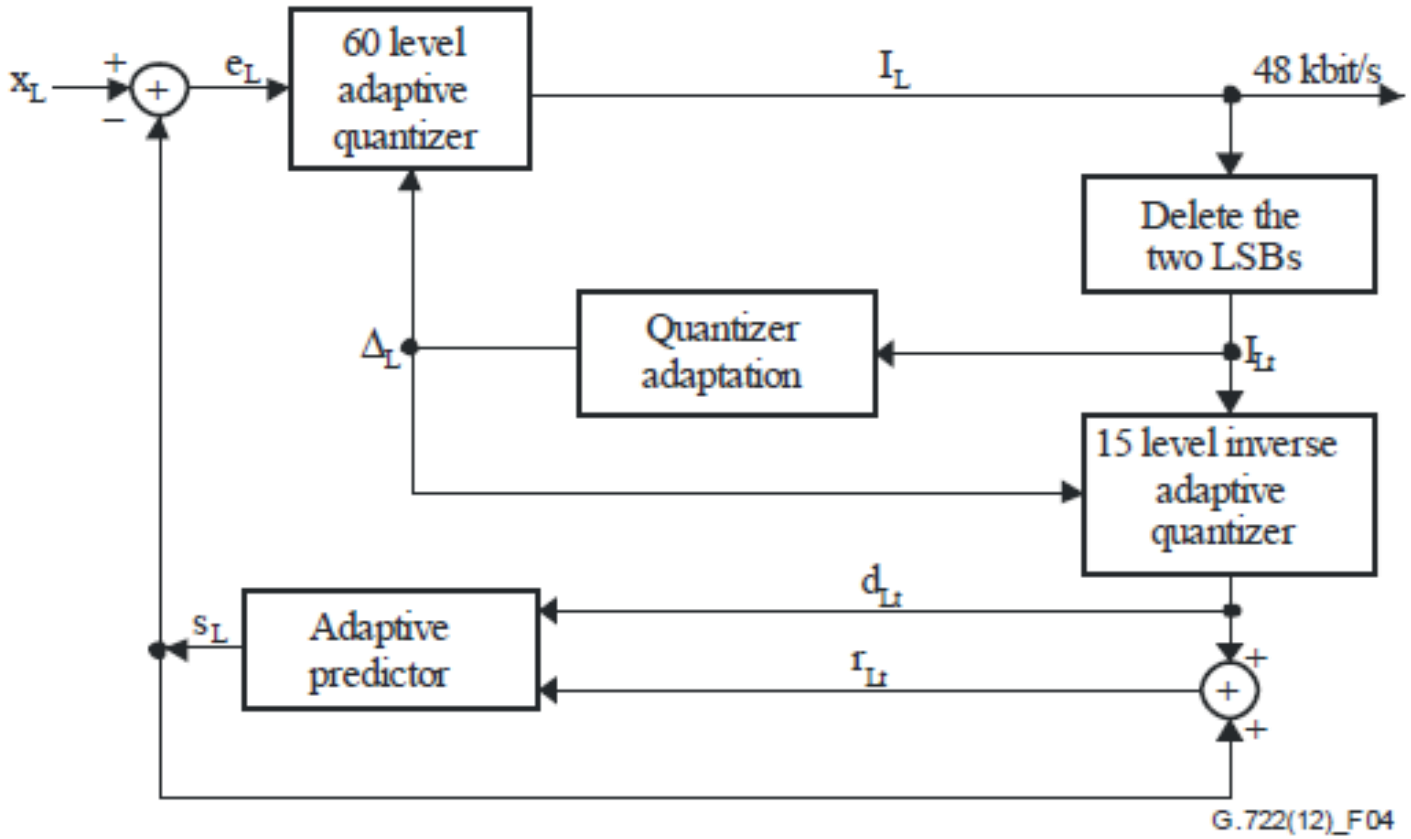


Figure 4 – Block diagram of the lower sub-band ADPCM encoder

G.722

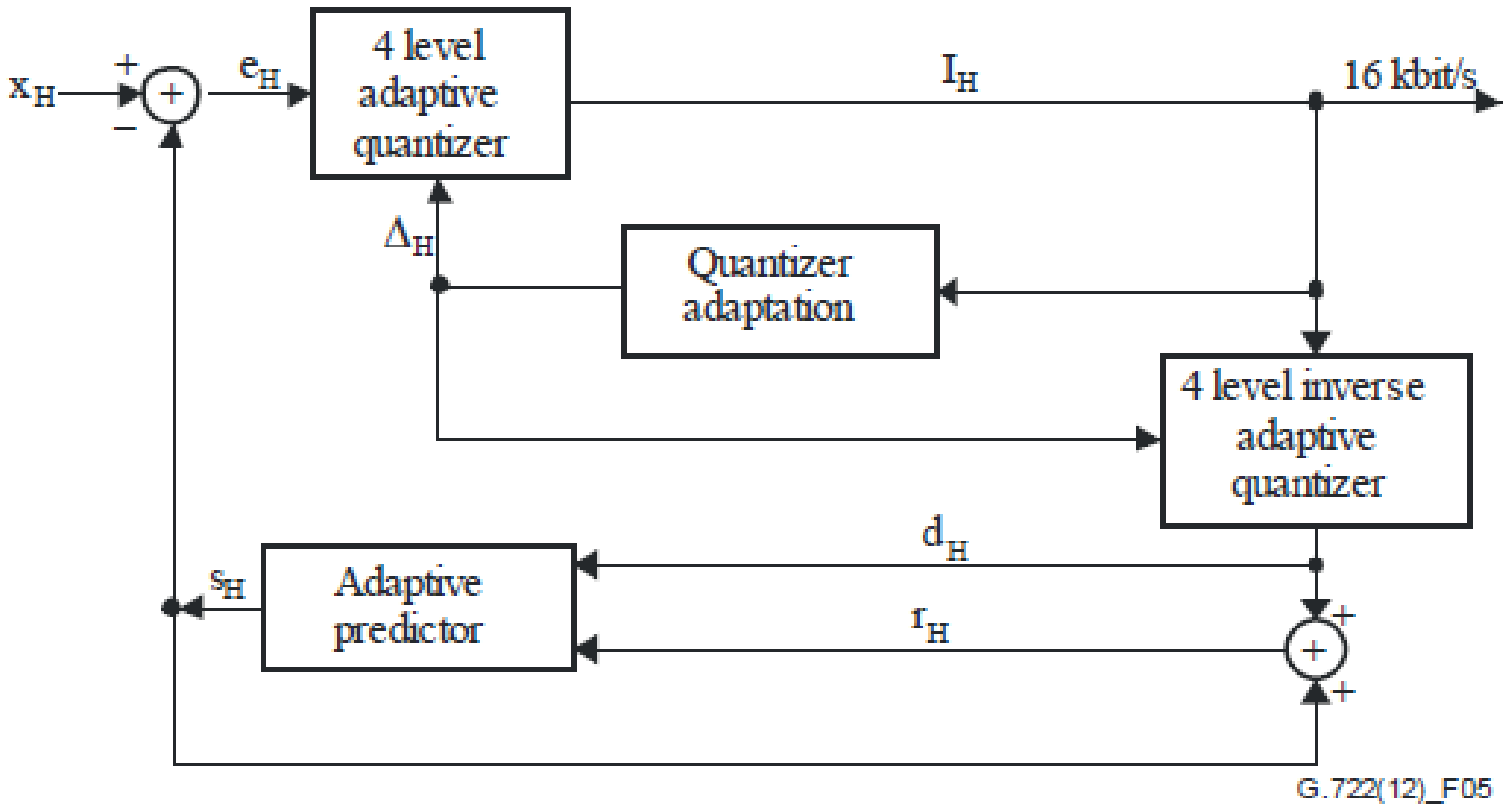


Figure 5 – Block diagram of the higher sub-band ADPCM encoder

G.722

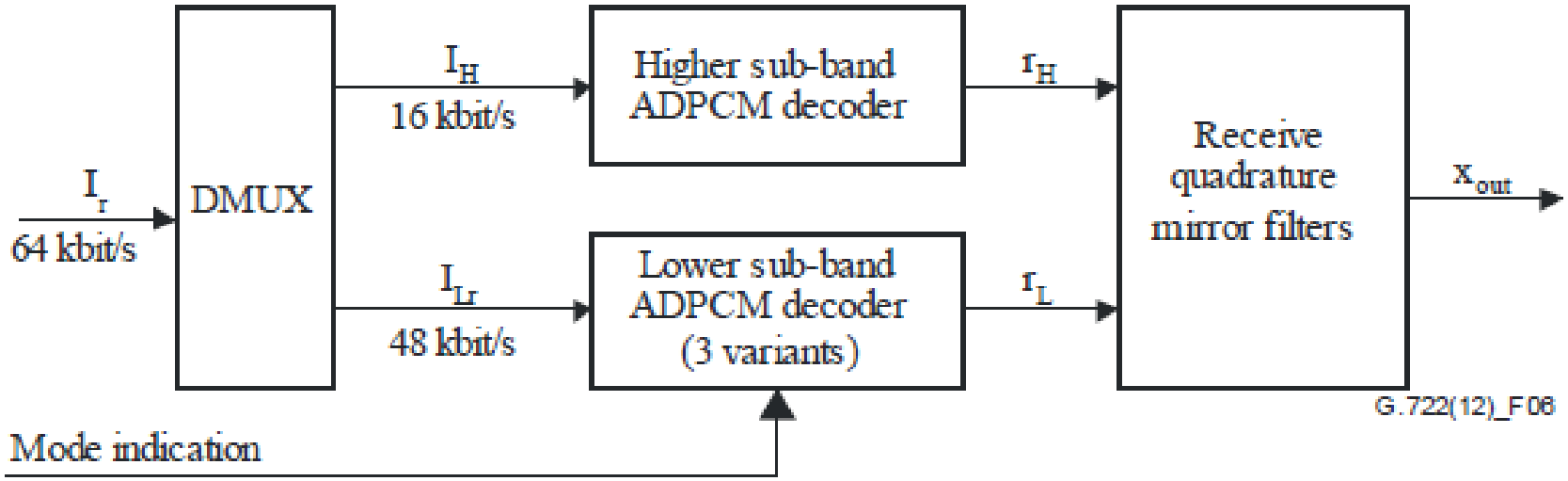


Figure 6 – Block diagram of the SB-ADPCM decoder

AMR (Adaptive Multi Rate)

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).

De forma similar a G.729, se basa en el modelo ACELP

- Ventanas de audio de 20 ms (160 muestras)
- 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)
- Por cada sub-ventana se obtienen los índices de los “codebooks” fijos y adaptivos y las ganancias.

AMR (Adaptive Multi Rate)

Según la forma en que se cuanticen los 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.

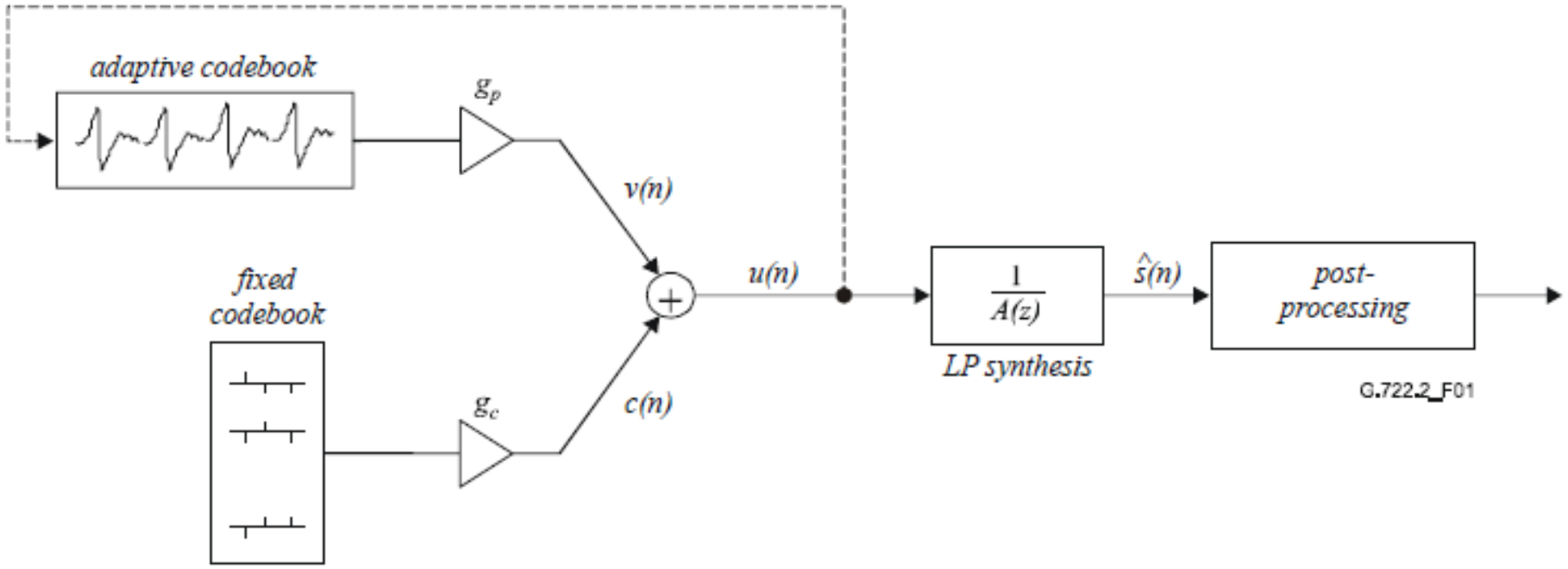
AMR-WB (G.722.2)

Codec de Banda Ancha (16 kHz), basado en un muestreo inicial de 14 bits por muestra

9 posibles velocidades entre 6.6 y 23.85 kb/s

Basado en CELP, utilizando un filtro de orden 16

AMR-WB (G.722.2)


adaptive codebook
g_p
$v(n)$
fixed codebook
g_c
$c(n)$
$u(n)$
$\frac{1}{A(z)}$
LP synthesis
$\hat{s}(n)$
post-processing
G.722.2_F01

SILK

Utilizado inicialmente por Skype.

Ancho de banda variable, entre 6 a 40 kb/s, trabajando entre las bandas angostas (8 kHz) y las bandas super anchas (superwideband) (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

OPUS

Soporta VBR (Variable Bit Rate) y CBR (Constant Bit Rate).

- El “default” es VBR

Utiliza “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

	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

EVS: Enhanced Voice Services

Diseñado para servicios de VoLTE (Voice over LTE)

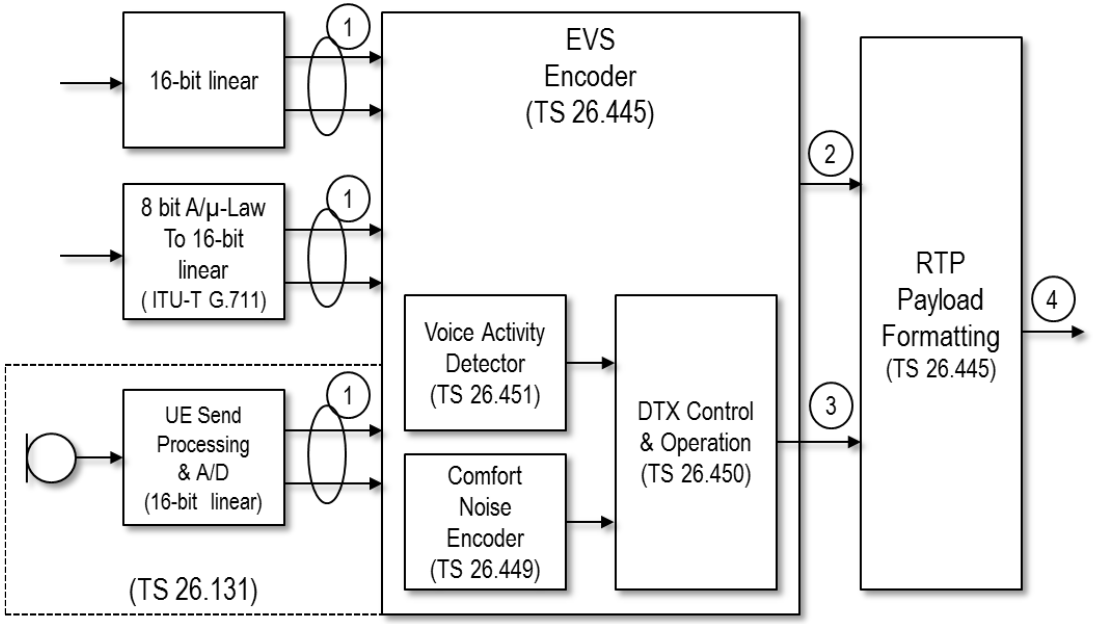
Es el primer códec desarrollado por 3GPP de banda completa (hasta 20 kHz)

Provee interoperabilidad con AMR-WB

Es de velocidad variable (VBR)

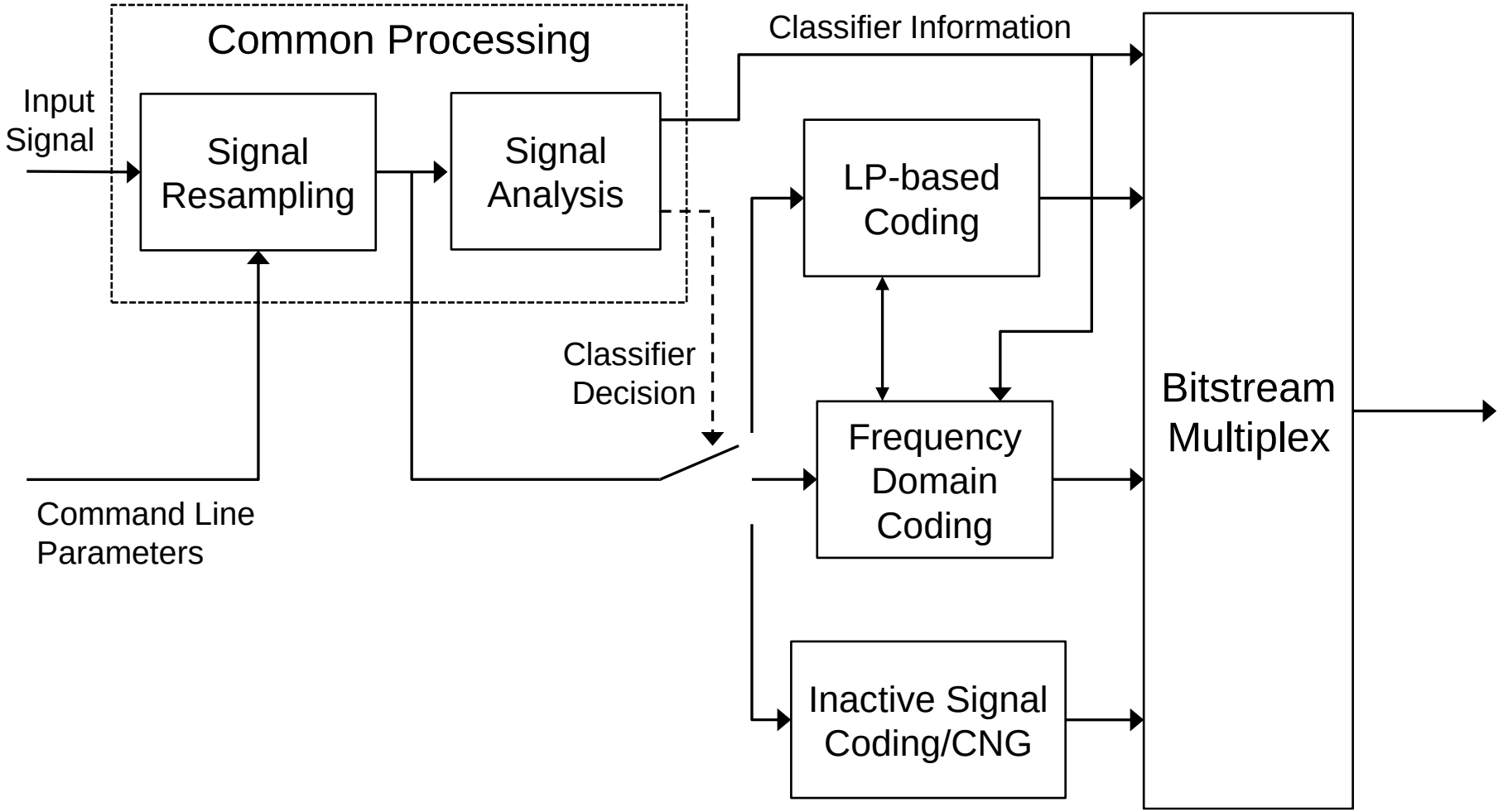
Bandwidth	Bit Rate (kbps)
Narrowband (NB)	5.9, 7.2, 8, 9.6, 13.2, 16.4, 24.4
Wideband (WB)	5.9, 7.2, 8, 9.6, 13.2, 16.4, 24.4, 32, 48, 64, 96, 128 (6.6 ~ 23.85 for AMR-WB IO)
Super-wideband (SWB)	9.6, 13.2, 16.4, 24.4, 32, 48, 64, 96, 128
Fullband (FB)	16.4, 24.4, 32, 48, 64, 96, 128

EVS: Enhanced Voice Services



1. 16-bit Linear PCM Samples and Sample Rate (8, 16, 32 or 48 kHz)
2. Encoded audio frame, 50 frames/s, number of bits/frame depending on the EVS codec mode
3. Encoded Silence Descriptor frames (variable frame rate)
4. RTP Payload Packets

EVS: Encoder


Common Processing
Input Signal
Signal Resampling
Signal Analysis
Classifier Information
LP-based Coding
Classifier Decision
Frequency Domain Coding
Bitstream Multiplex
Command Line Parameters
Inactive Signal Coding/CNG

SATIN

Codec desarrollado por Microsoft en 2020

Pensado para funcionar con buena calidad en condiciones de poco ancho de banda y alto porcentaje de pérdida de paquetes

Basado en tecnologías de “Inteligencia Artificial”

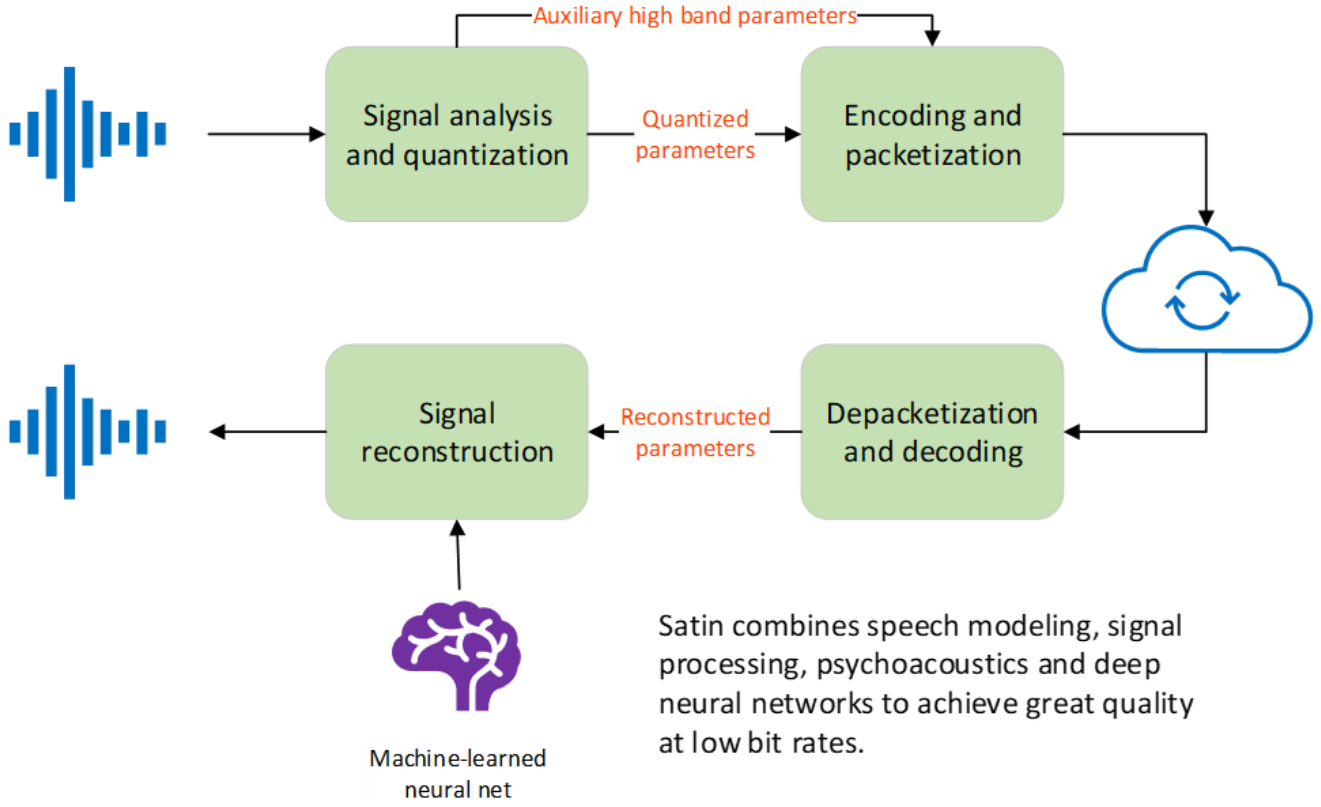
- Para reducir la tasa de bits requerida, Satin solo codifica y transmite ciertos parámetros en las bandas de frecuencia más bajas. En el decodificador, Satin utiliza redes neuronales profundas para estimar los parámetros de banda alta a partir de los parámetros de banda baja recibidos.

Codifica cada paquete de forma independiente, por lo que el efecto de perder un paquete no afecta la calidad de los paquetes posteriores.

Basado en: https://techcommunity.microsoft.com/t5/microsoft-teams-blog/satin-microsoft-s-latest-ai-powered-audio-codec-for-real-time/ba-p/2141382

SATIN

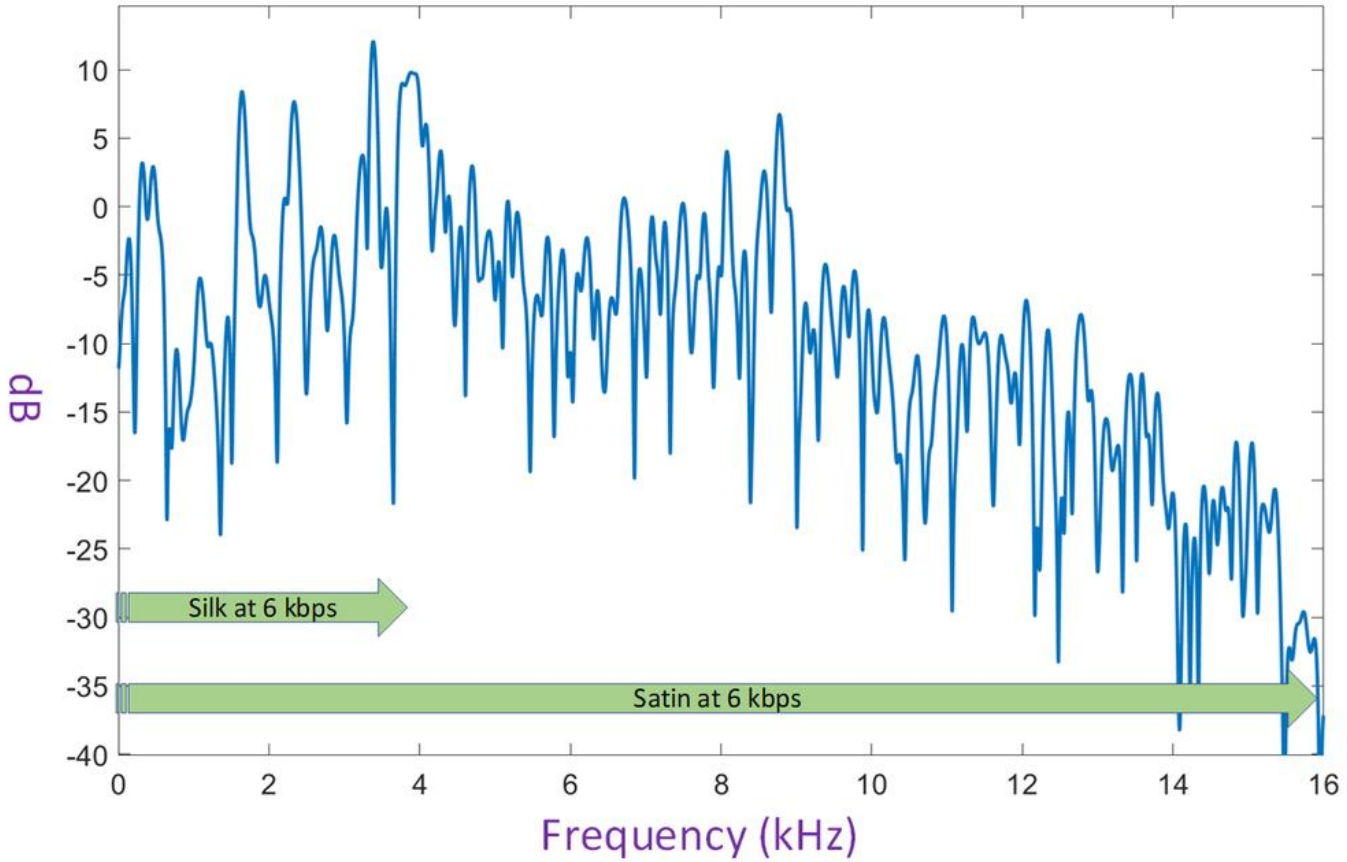
Para reducir la tasa de bits requerida, Satin solo codifica y transmite ciertos parámetros en las bandas de frecuencia más bajas. En el decodificador, Satin utiliza redes neuronales profundas para estimar los parámetros de banda alta a partir de los parámetros de banda baja recibidos



Tomado de: https://techcommunity.microsoft.com/t5/microsoft-teams-blog/satin-microsoft-s-latest-ai-powered-audio-codec-for-real-time/ba-p/2141382

SATIN

“Componentes de frecuencia del sonido /t/ en la palabra “suit". Hay una cantidad significativa de energía mucho más allá del límite de banda angosta de 4 kHz e incluso del límite de banda ancha de 8 kHz. La conservación de la energía en los componentes espectrales superiores da como resultado un sonido mucho más natural.”



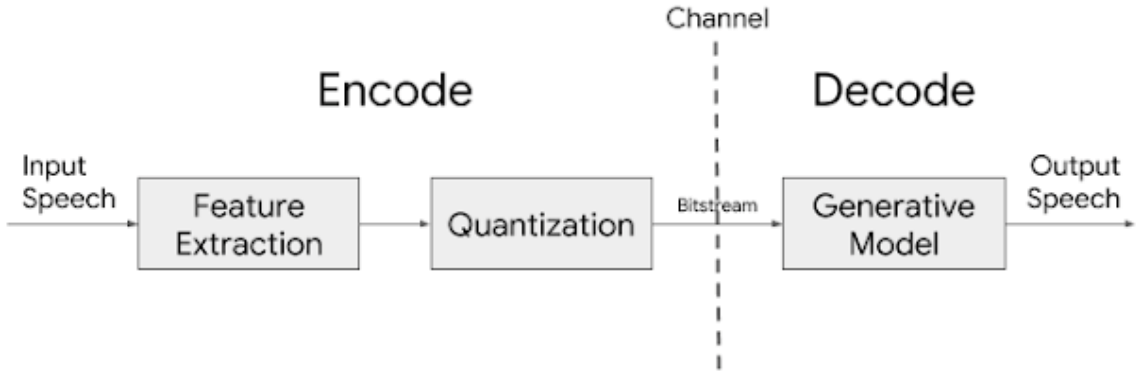
Tomado de: https://techcommunity.microsoft.com/t5/microsoft-teams-blog/satin-microsoft-s-latest-ai-powered-audio-codec-for-real-time/ba-p/2141382

LYRA

Codec desarrollado por Google en 2021[1]

Utiliza ventanas de 40 ms, y representa el audio con “espectrogramas de Mel” logarítmicos (representación de la energía en diferentes bandas de frecuencia).

En la decodificación utiliza un “modelo generativo” basado en “WaveNetEQ”, mejorado con redes neuronales de latencia 90 ms y bajos requerimientos de procesamiento



Basado en: https://ai.googleblog.com/2021/02/lyra-new-very-low-bitrate-codec-for.html

[1] KLEIJN, W. Bastiaan, et al. **“Generative Speech Coding with Predictive Variance Regularization”**. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021. p. 6478-6482.

IVAS: Immersive Voice and Audio Services
(en desarrollo)

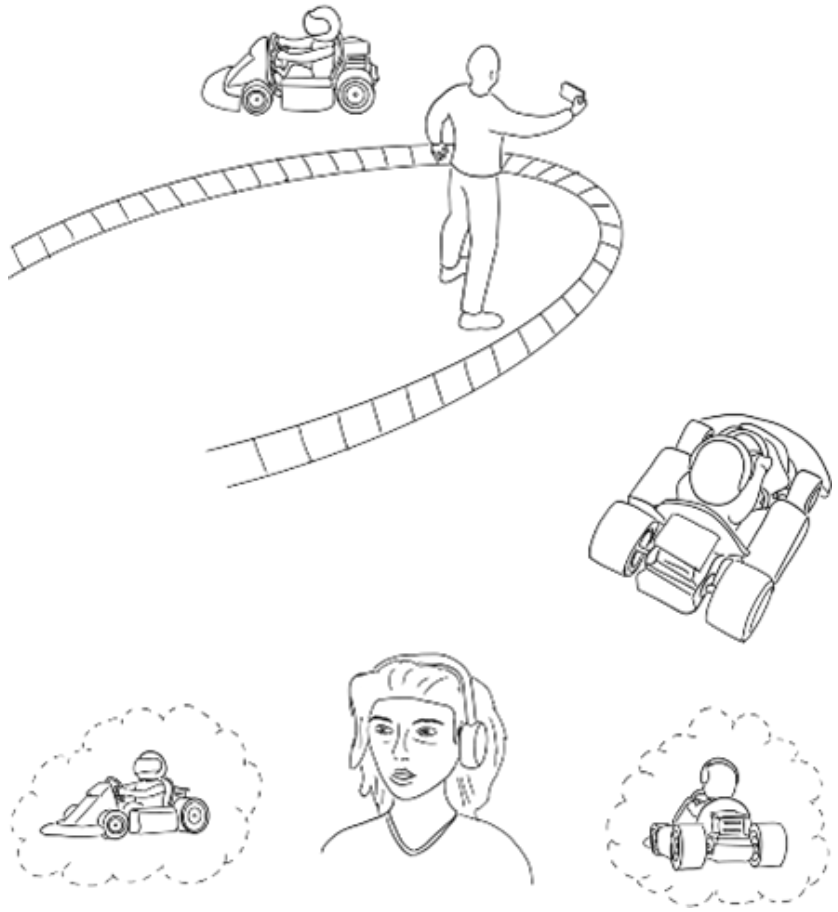
Proyecto Immersive Voice and Audio Services (IVAS) de 3GPP

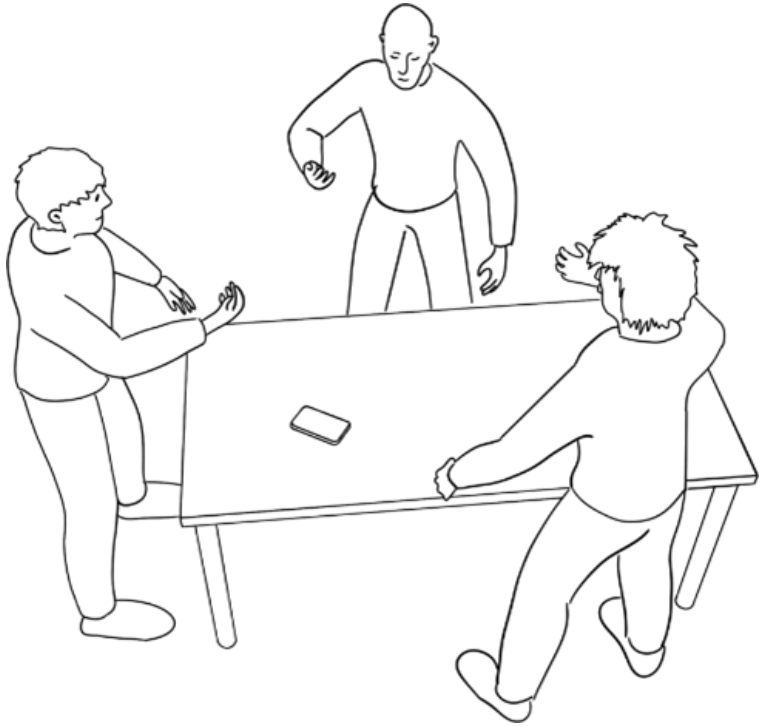
El objetivo general es desarrollar un único códec de audio de propósito general para ***servicios y aplicaciones inmersivas.***

Una experiencia audiovisual *inmersiva* implica, para el componente de audio, que una impresión de sonido espacial es convincentemente coherente con la escena visual presentada. Además, el usuario debe poder moverse, dentro de ciertos límites definidos por la aplicación, a lo largo de la escena, y el componente de audio se ajustará para reflejar la orientación / posición espacial del usuario.

El objetivo es que este nuevo codec pueda funcionar sobre redes 5G.

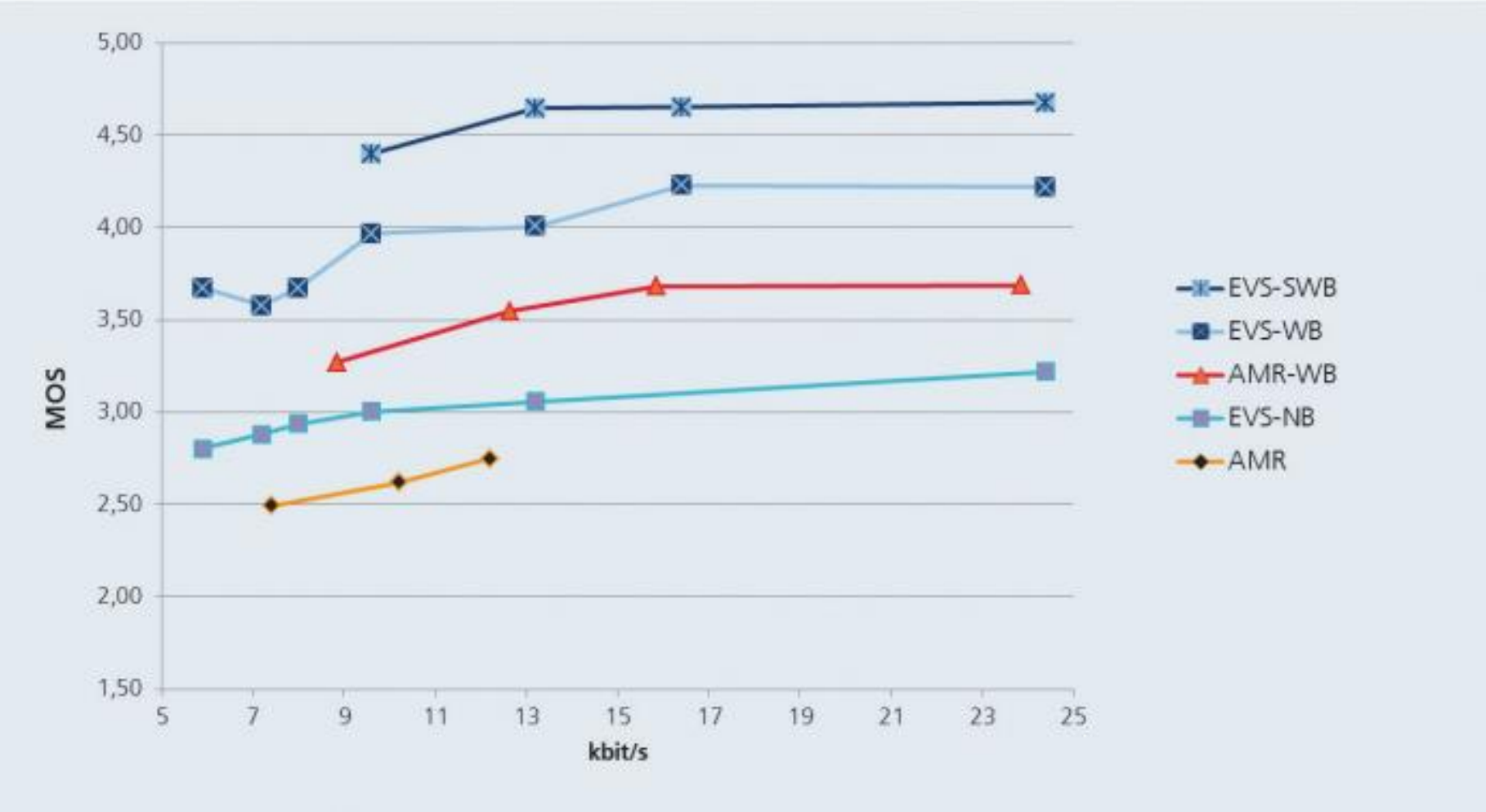
IVAS: Immersive Voice and Audio Services
(en desarrollo)





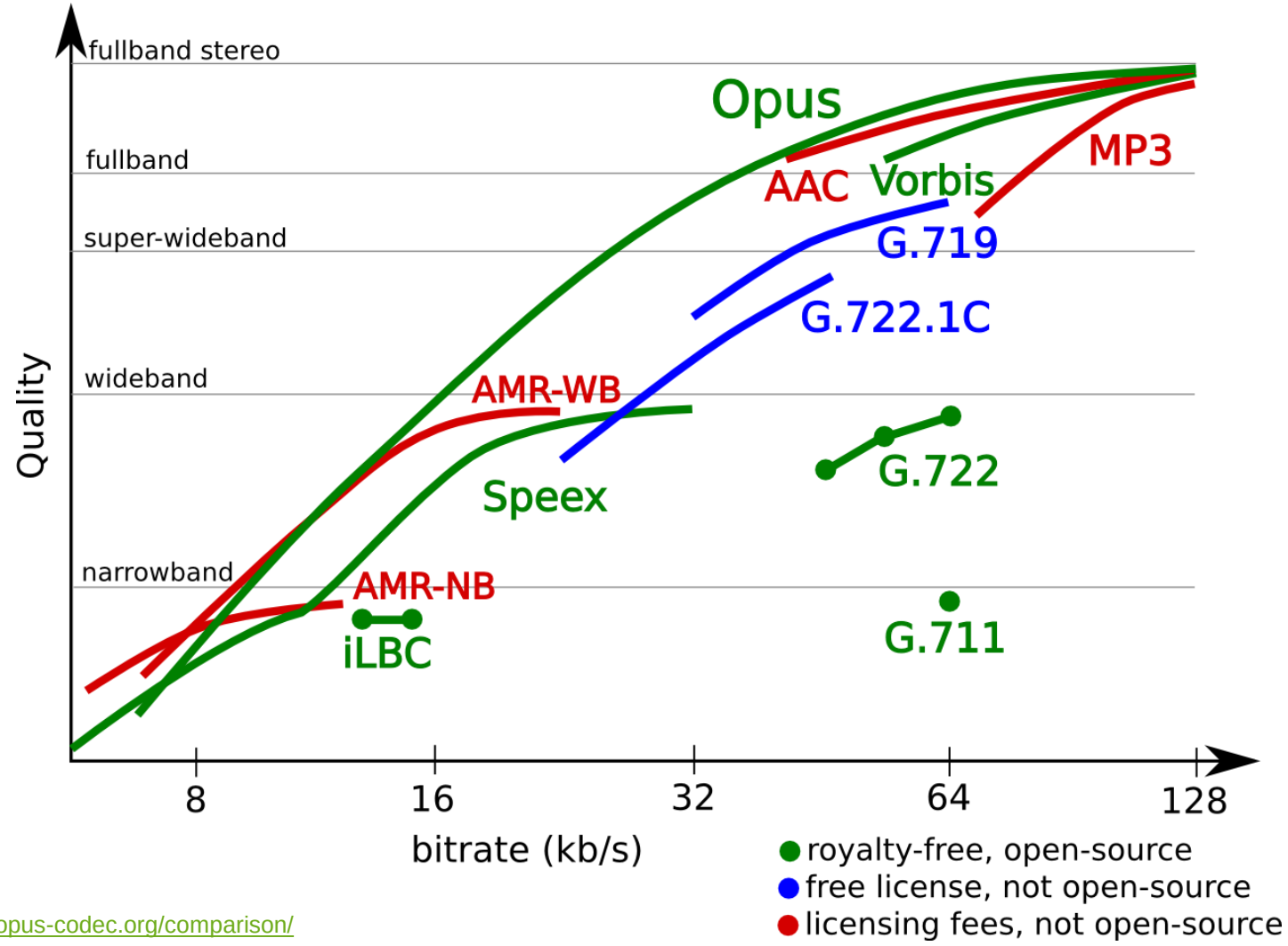
Tomado de https://www.3gpp.org/technologies/ivas-highlights

Comparación de codecs de voz



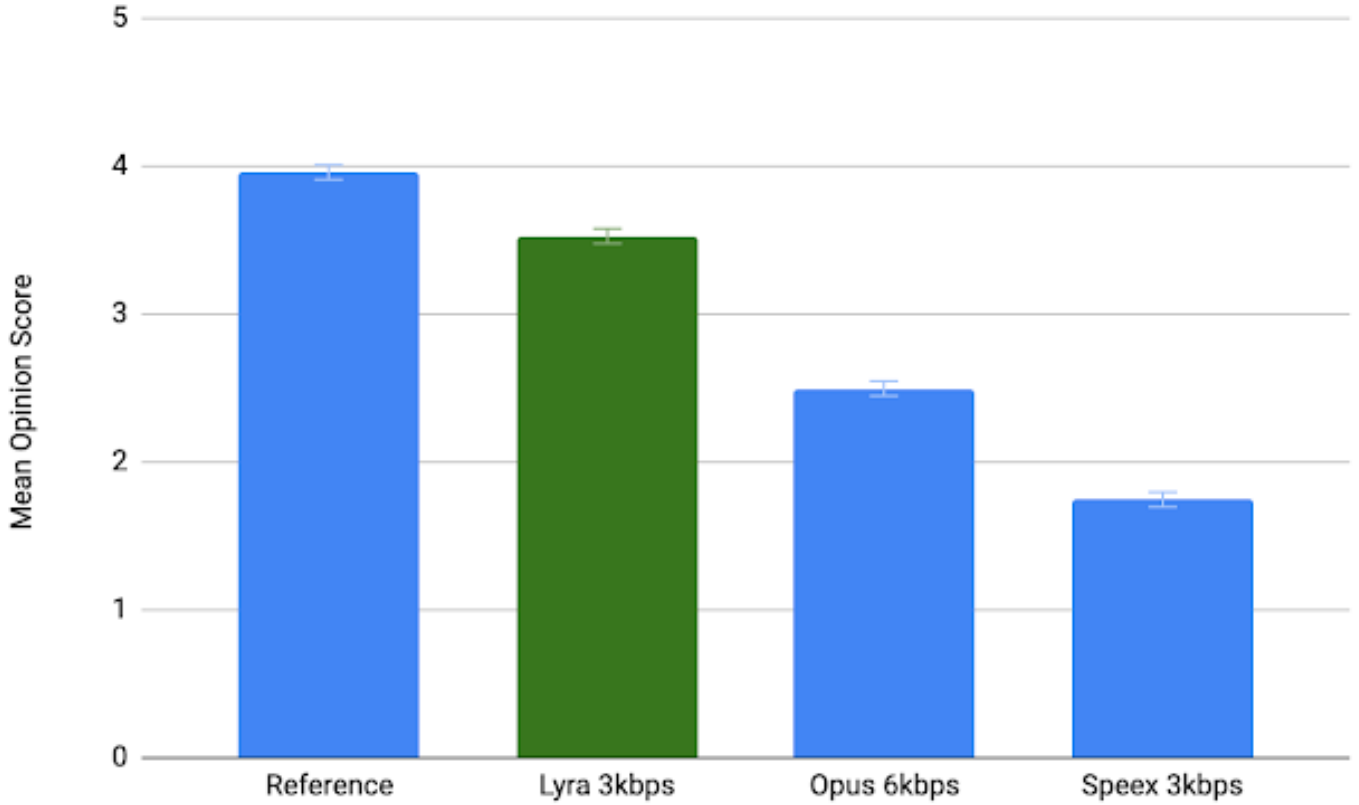
Tomado de The Future of Communication: Full-HD Voice powered by EVS and the AAC-ELD Family, Fraunhofer

Comparación de codecs de voz



Tomado de http://www.opus-codec.org/comparison/

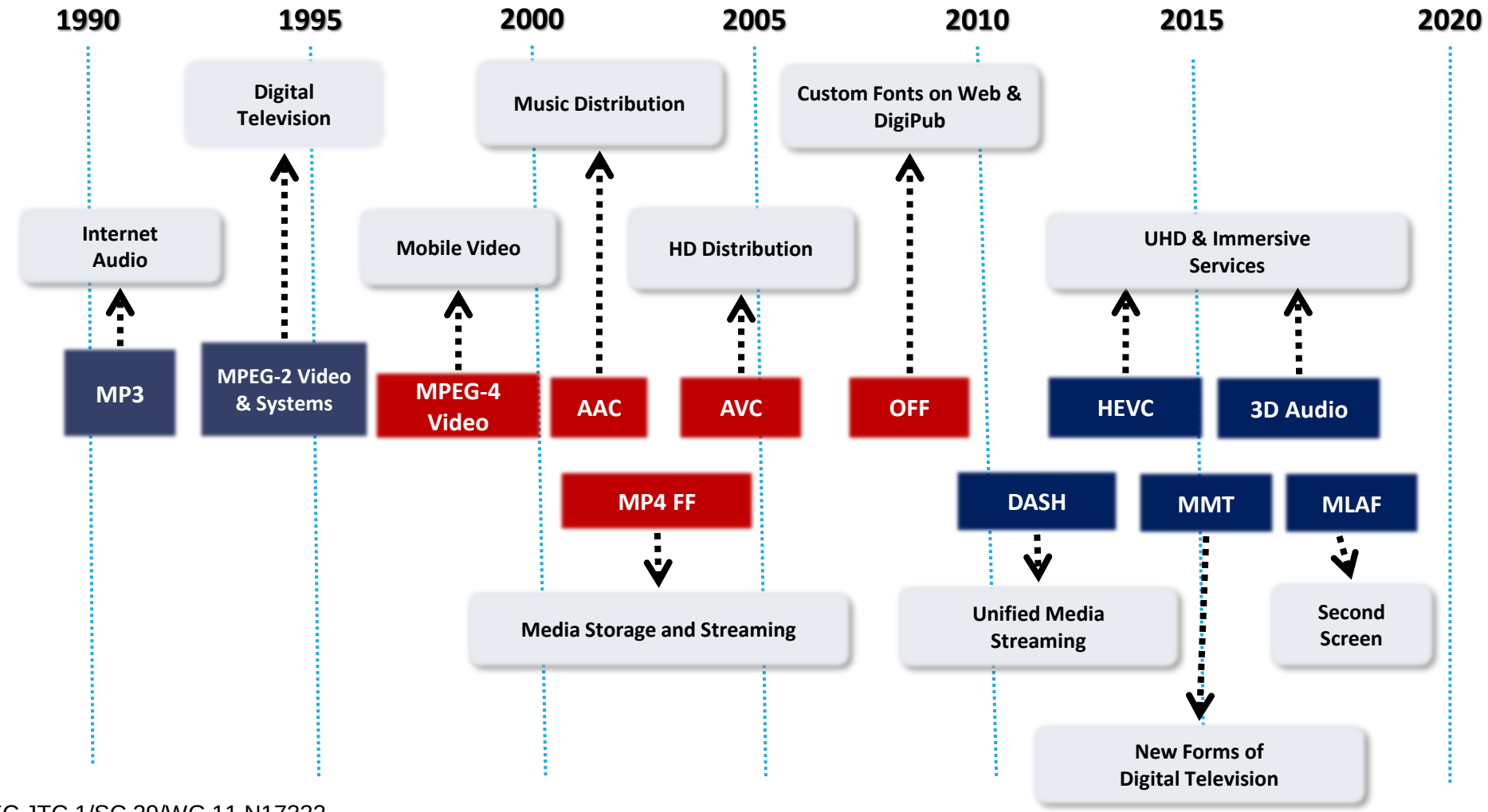
Comparación de codecs de voz



Tomado de : https://ai.googleblog.com/2021/02/lyra-new-very-low-bitrate-codec-for.html

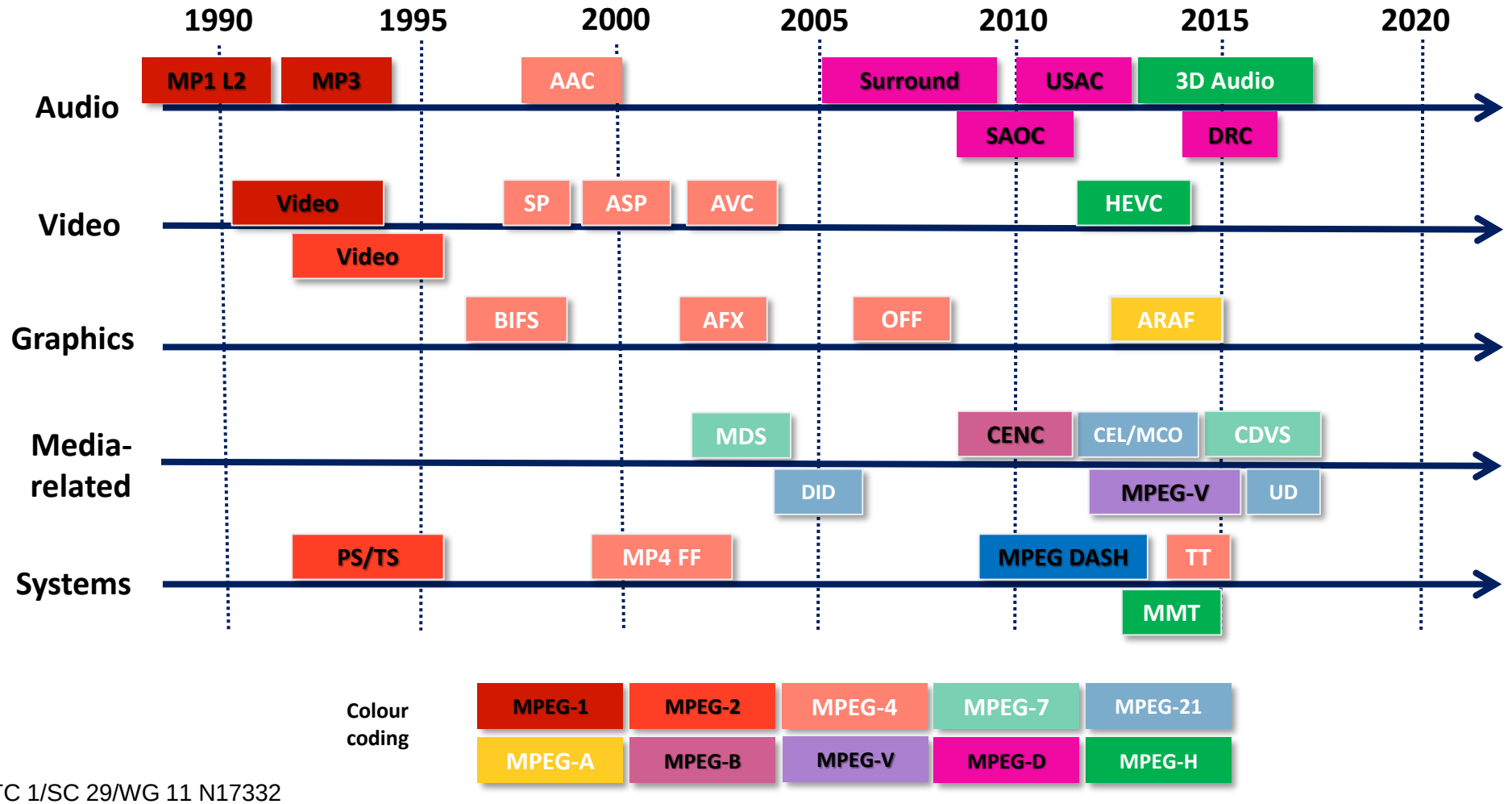
Codificación de Audio

Evolución de los estándares MPEG



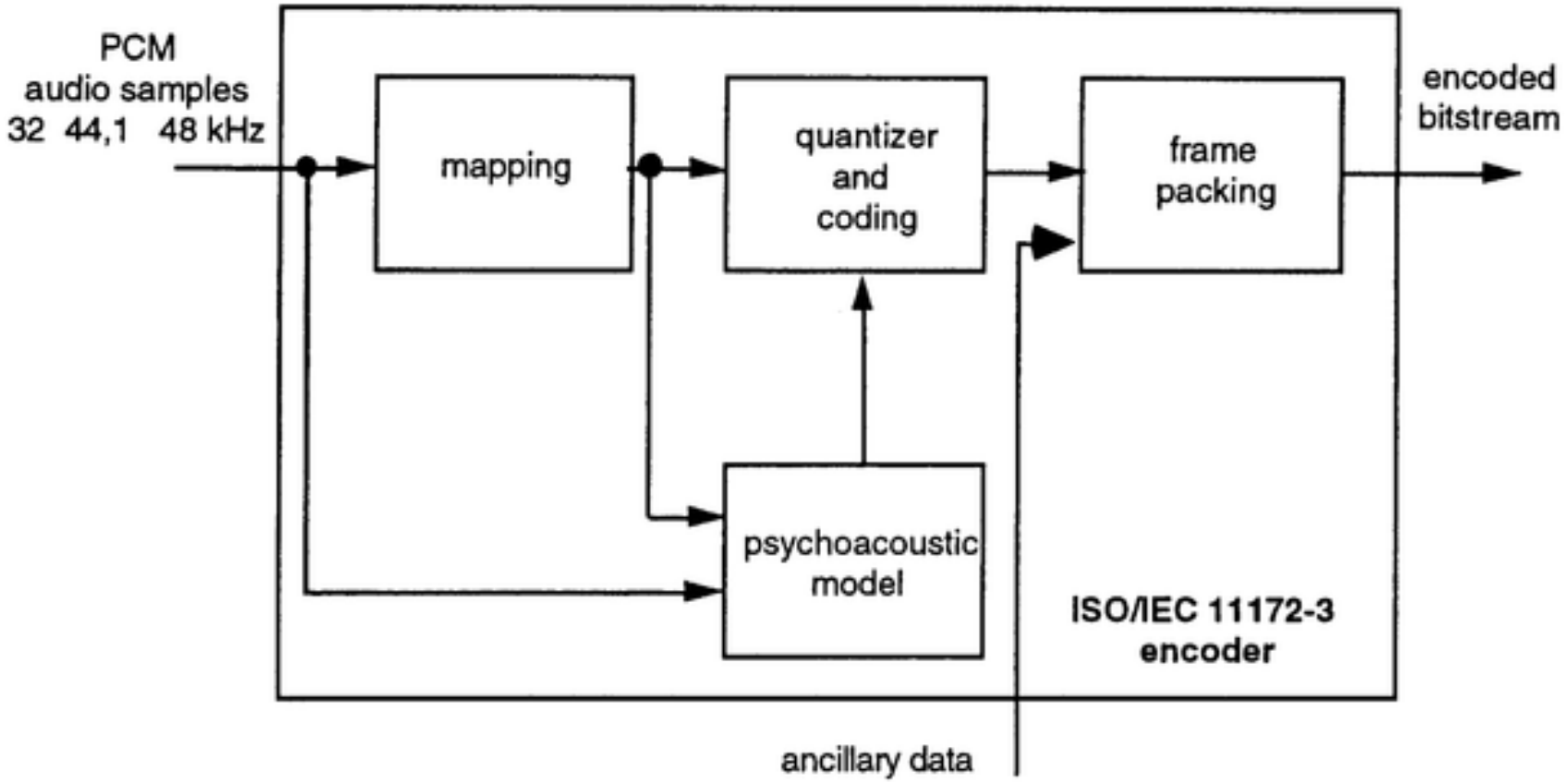
Basado en: ISO/IEC JTC 1/SC 29/WG 11 N17332

Evolución de los estándares MPEG

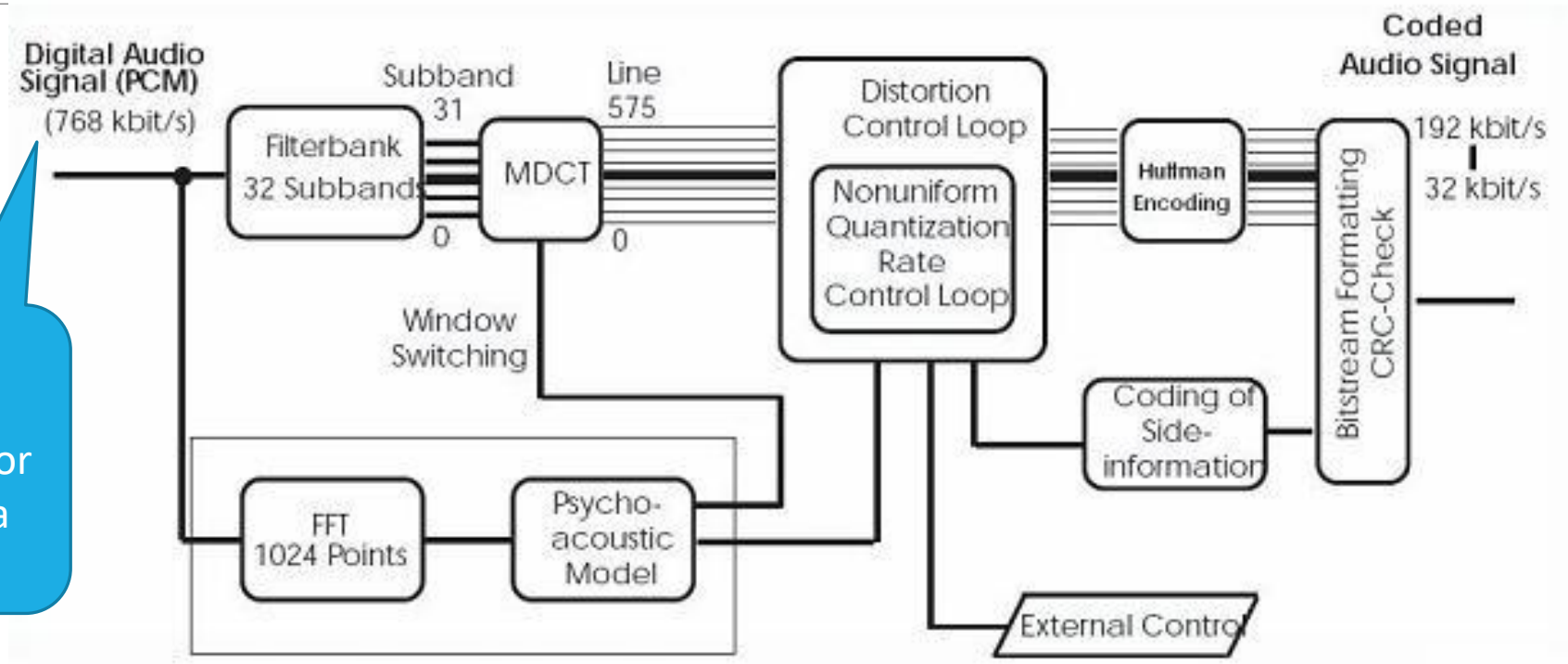


Basado en: ISO/IEC JTC 1/SC 29/WG 11 N17332

MP3 - encoder


PCM
audio samples
32 44,1 48 kHz
mapping
quantizer
and
coding
frame
packing
encoded
bitstream
psychoacoustic
model
ISO/IEC 11172-3
encoder
ancillary data

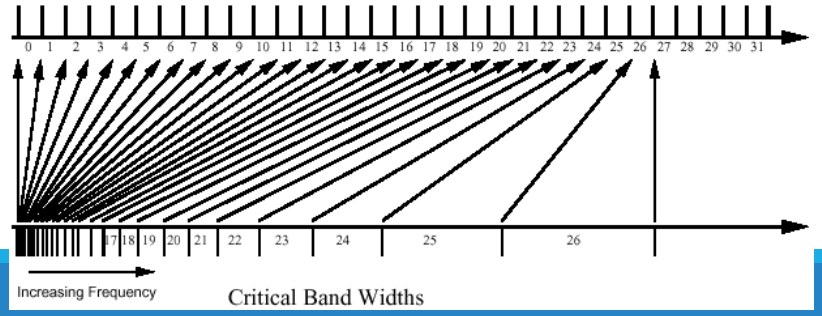
MP3 - encoder



48 kHz,
16 bits por
muestra

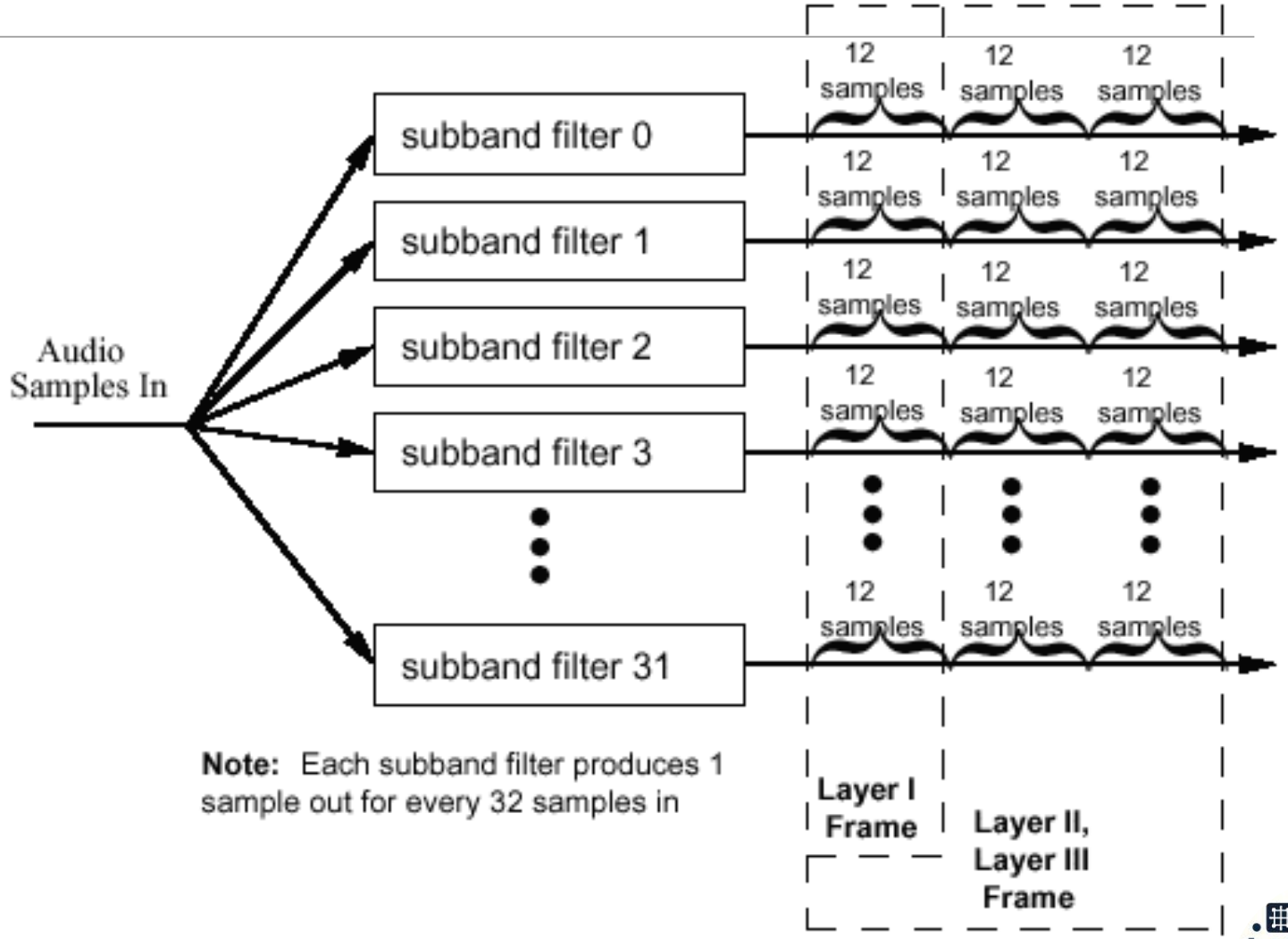
Tomado de
http://wiki.sj.ifsc.edu.br/wiki/index.php/MP3_(Artigo_Completo)

MPEG/Audio Filter Bank Bands



Tomado de https://cs.stanford.edu/people/eroberts/courses/soco/projects/data-compression/lossy/mp3/hybrid_filter.htm

MP3 - encoder

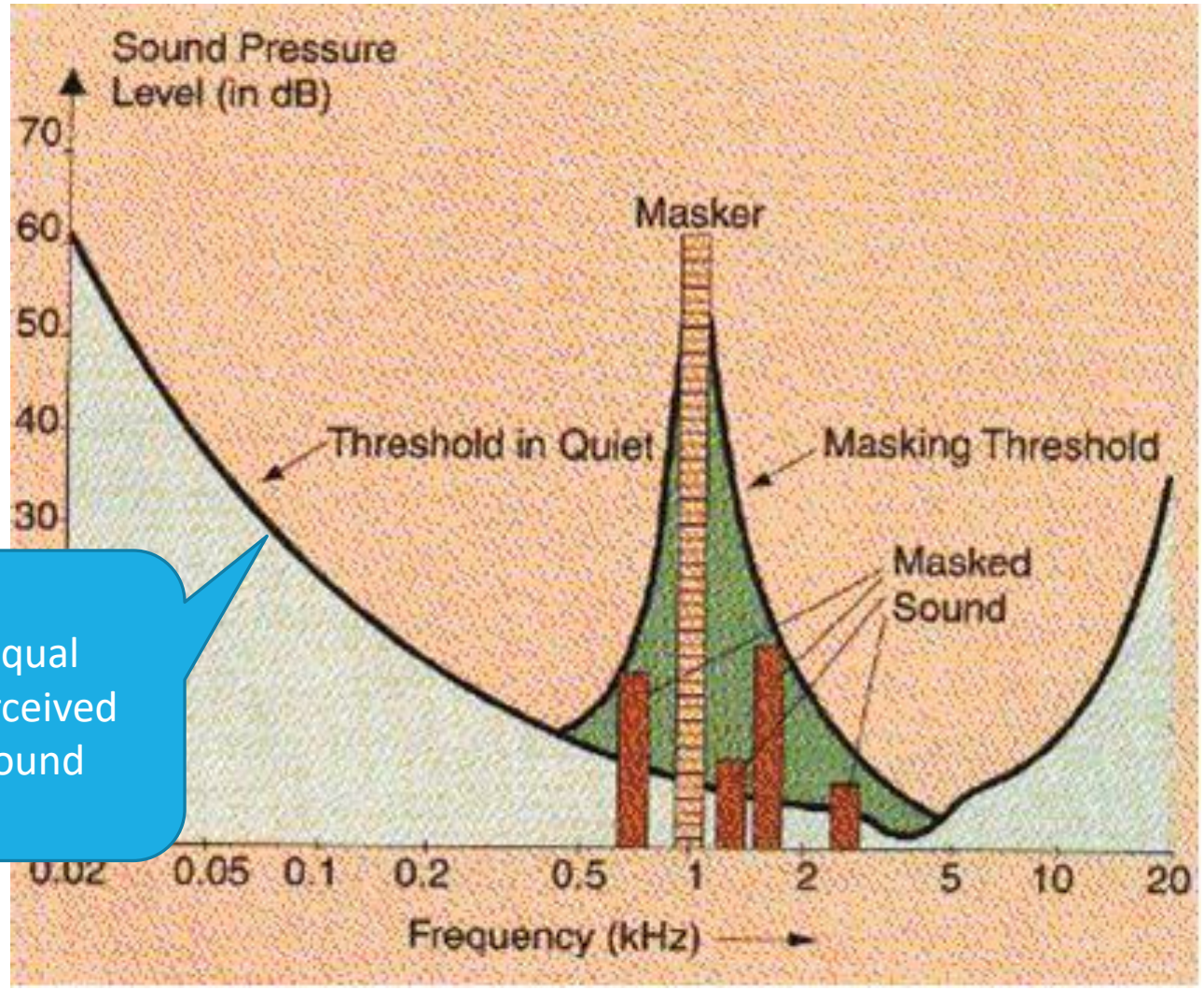


Using MPEG compression techniques, data can be reduced to the following percentages while still maintaining CD sound quality:

25% By Layer 1
16% to 12% By Layer 2
10% to 8% By Layer 3

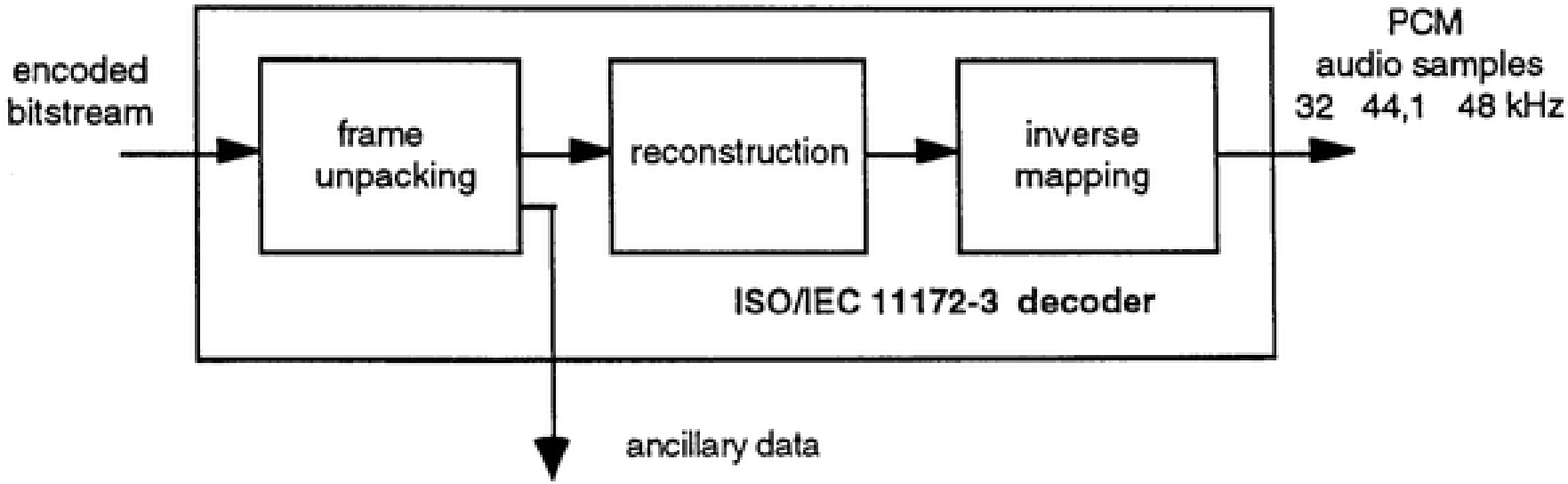
Tomado de
https://cs.stanford.edu/people/eroberts/courses/soco/projects/data-compression/lossy/mp3/layer_coding.htm

MP3- uso de “enmascaramiento”



Equal perceived sound

MP3 - decoder


encoded
bitstream
frame
unpacking
reconstruction
inverse
mapping
PCM
audio samples
32 44,1 48 kHz
ISO/IEC 11172-3 decoder
ancillary data

MP3

Se pueden utilizar diferentes capas, con complejidad y rendimiento del codificador crecientes

Un decodificador de capa N es capaz de decodificar datos que se han codificado en la capa N y todas las capas inferiores.

Capa I

- Contiene el mapeo básico de la entrada de audio digital, segmentado en 32 sub-bandas, aplicando un modelo “psicoacústico” para determinar una asignación adaptativa de bits y escalas de cuantificación en cada sub-banda. El retardo de codificación / decodificación mínimo teórico para la Capa I es de aproximadamente 19 ms.

Capa II

- Proporciona codificación adicional. El retardo de codificación / decodificación mínimo teórico para la capa II es de aproximadamente 35 ms.

Capa III

- Introduce mayor resolución de frecuencia basada en un banco de filtros híbrido, en dos etapas. Añade un cuantificador (no uniforme) diferente, segmentación adaptativa y codificación entrópica de los valores cuantificados. El retardo de codificación / decodificación mínimo teórico para la capa III es de aproximadamente 59 ms.

La codificación estéreo se puede añadir como una característica adicional a cualquiera de las capas.

AAC: Advanced Audio Coding

Parte de MPEG 4 – audio (1997)

48 full band audio channels

Entrada de 8 a 96 kHz

Usa MDCT (Modified Discrete Cosine Transform) en el “banco de filtros”

MPEG 4- audio part 3

- CELP (Code Excited Linear Prediction, para voz)
- AAC (Advanced Audio Coding)
- MIDI (Musical Instrument Digital Interface)
- Part 14 – mp4

El formato AAC permite incluir legalmente la protección de los derechos de autor. Aquellos archivos de audio sin autorización, que tengan protección anticopia, no funcionarán en AAC

TNS: Temporal Noise Shaping

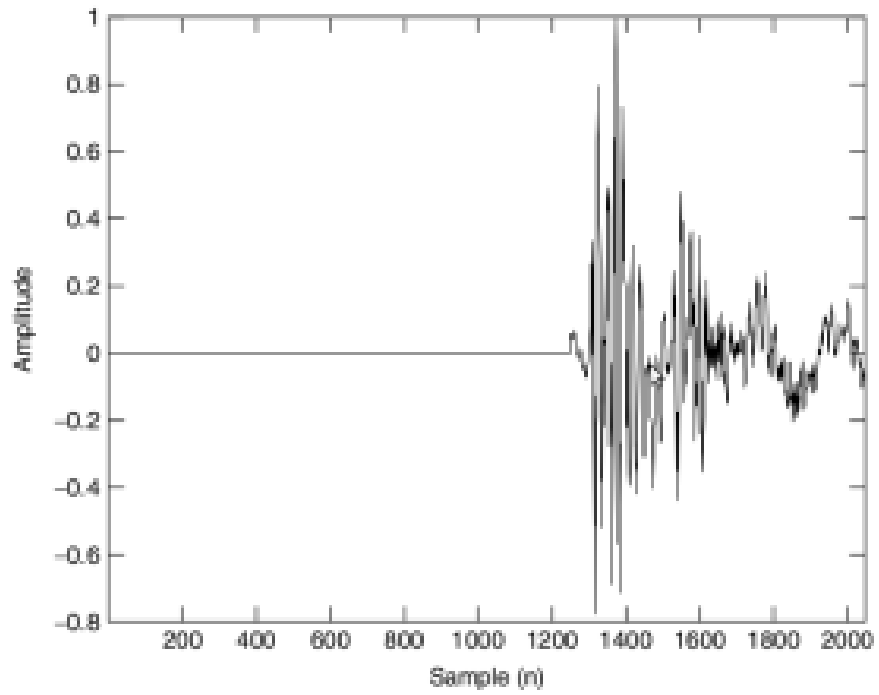
Fue introducido por primera vez en la versión de AAC de MPEG-2.

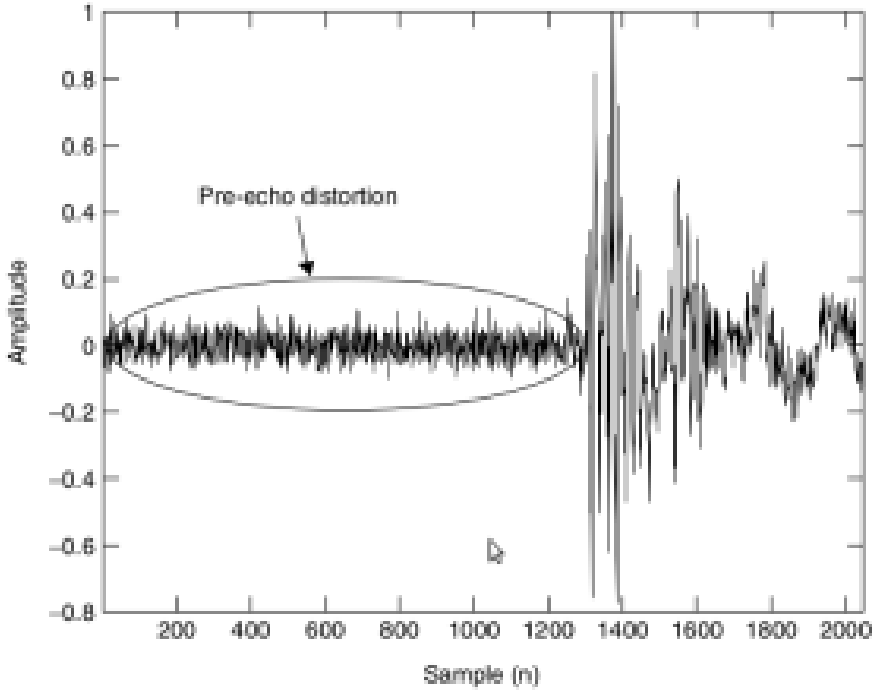
Cuando se tiene una señal con mucho “ataque”, luego de una región de poca energía y que comienza sobre el final de un bloque, aparece un fenómeno que se denomina “distorsión de pre-eco”

El error de cuantización introducido en frecuencia se esparce en el tiempo generando ruido no enmascarado durante la región de poca energía.

El bloque Temporal Noise Shaping soluciona este problema.

TNS: Temporal Noise Shaping





PNS: Perceptual Noise Substitution

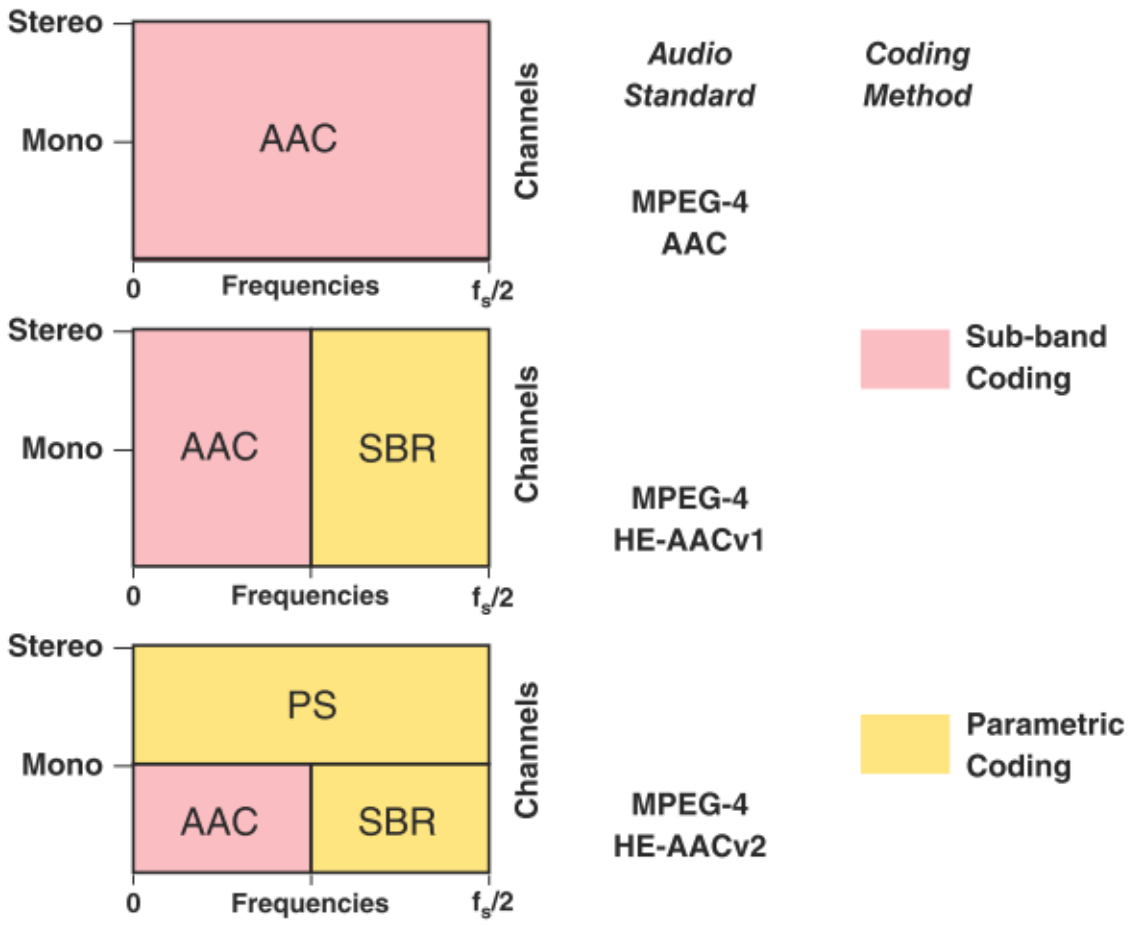
Basado en el concepto de que “todo el ruido suena parecido”.

Se detectan los coeficientes espectrales en forma de ruido y no se codifican.

Lo que se transmite es una indicación junto a la potencia de estos coeficientes, para cada banda de frecuencia.

El decodificador reemplaza estos coeficientes por vectores pseudo aleatorios con la potencia correspondiente.

AACplus o High Efficency (HE-AAC)



Tomado de MPEG-2 SYSTEMS Fundamentals and Evolution of Paving the MPEG Road, Jan van der Meer, Wiley, 2014

Spectral Band Replication (SBR)

Se “estima” la banda alta de frecuencias en función de la banda baja

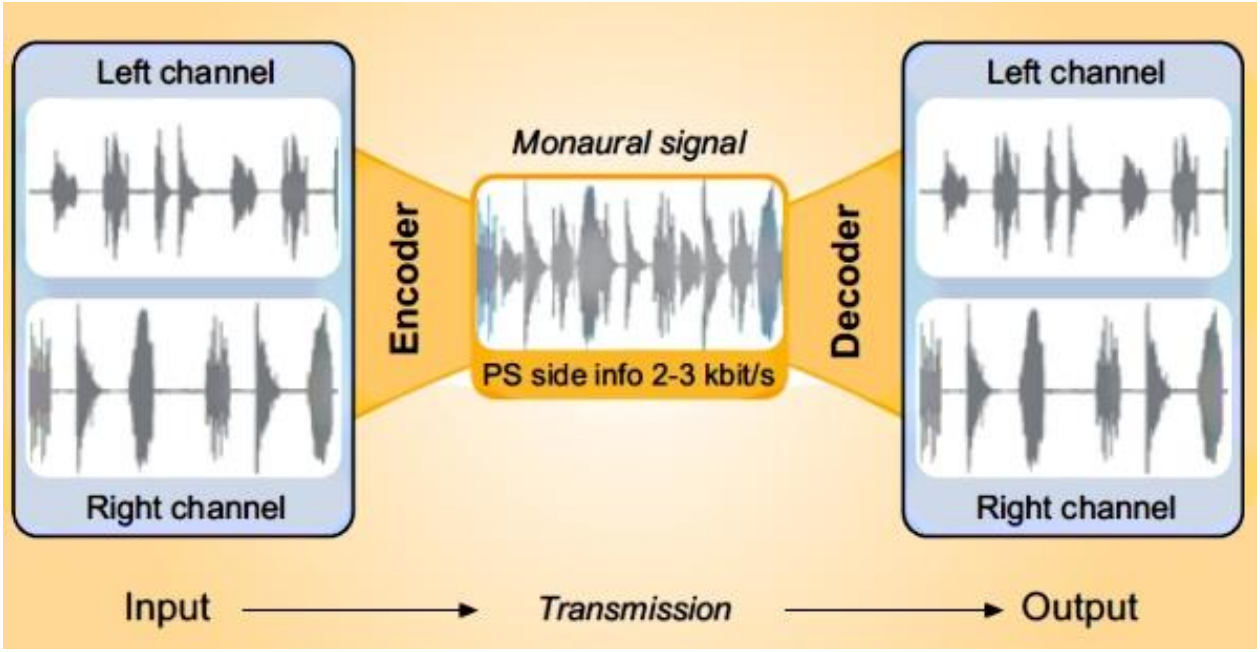
Se basa en transponer la banda baja (hasta 4-12 kHz) hacia la alta, y codificar eficientemente ciertos parámetros que permitan reconstruirla

Los decodificadores que no soportan SBR pueden decodificar los datos, pero obtendrán una señal de salida en banda limitada

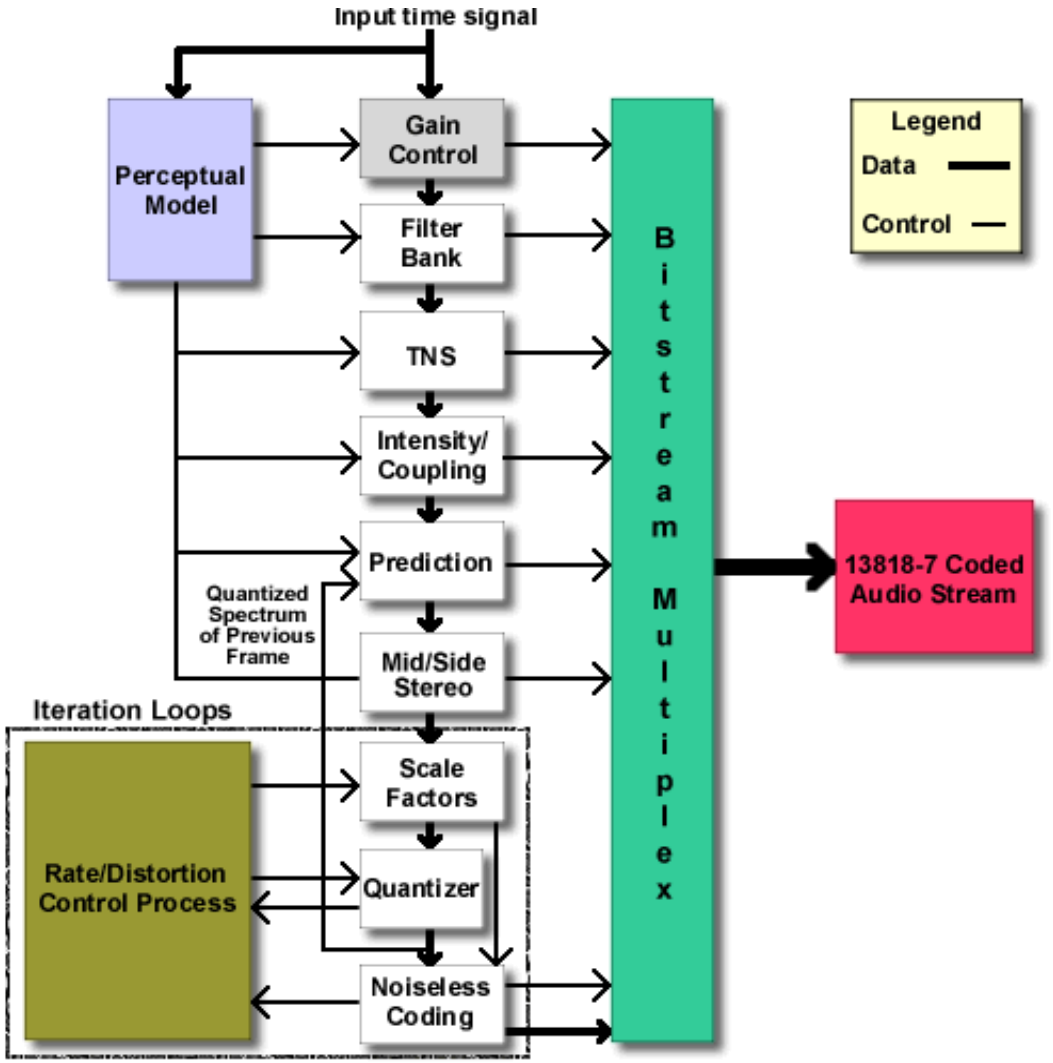
Parametric Stereo (PS)

En vez de realizar una transmisión estéreo, se extraen parámetros de cada uno de los canales y se hace un *downmix* a una señal “mono”

La señal mono resultante es codificada con HE-AAC y luego transmitida, conjuntamente con los datos paramétricos de “PS”



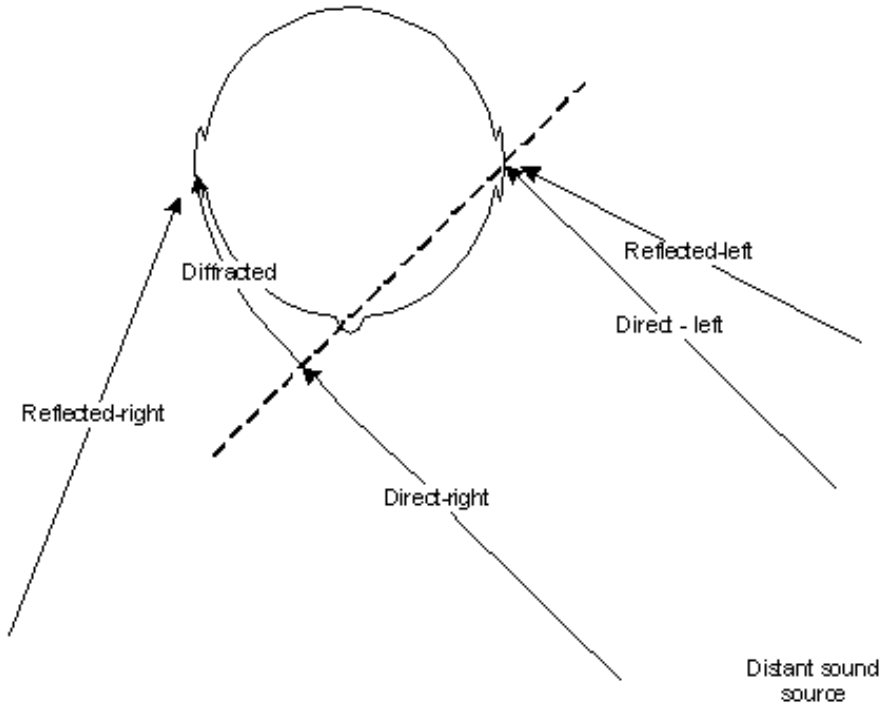
AAC


Input time signal
Perceptual Model
Gain Control
Filter Bank
TNS
Intensity/ Coupling
Prediction
Mid/Side Stereo
Quantized Spectrum of Previous Frame
Iteration Loops
Rate/Distortion Control Process
Scale Factors
Quantizer
Noiseless Coding
Bitstream Multiplex
Legend
Data
Control
13818-7 Coded Audio Stream

MPEG – D Surround Audio Coding (SAC)

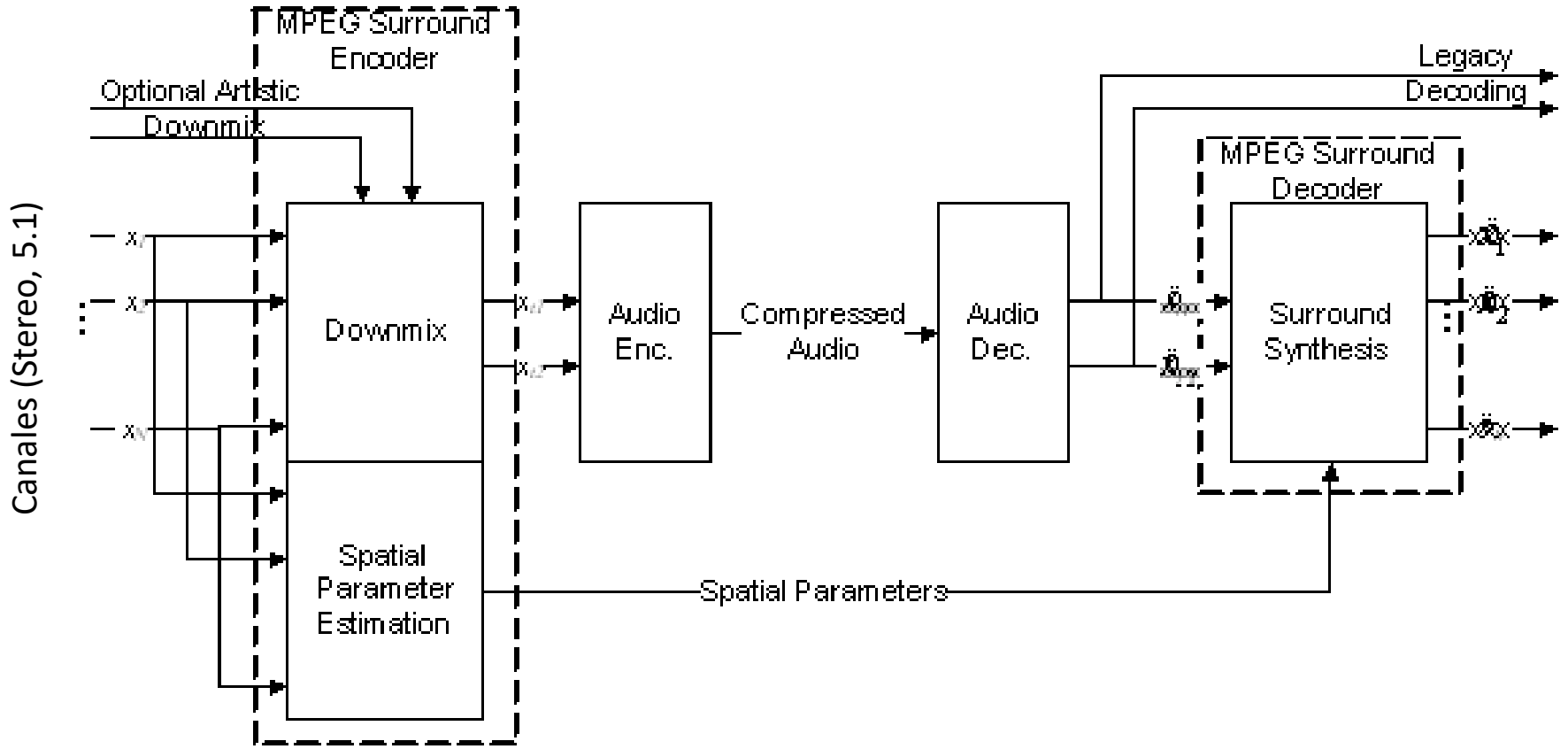
Explota nuestra capacidad de percibir el sonido en tres dimensiones y captura esa percepción en un conjunto compacto de parámetros:

- Diferencias de nivel interaural (ILD)
- Diferencias de tiempo interaural (ITD)
- Coherencia interaural (IC)



Tomado de: https://mpeg.chiariglione.org/standards/mpeg-d/mpeg-surround

MPEG – D Surround Audio Coding (SAC)

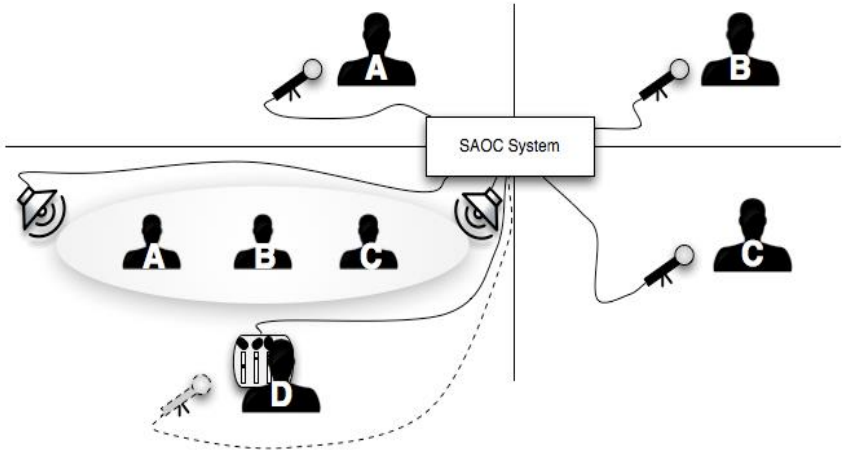


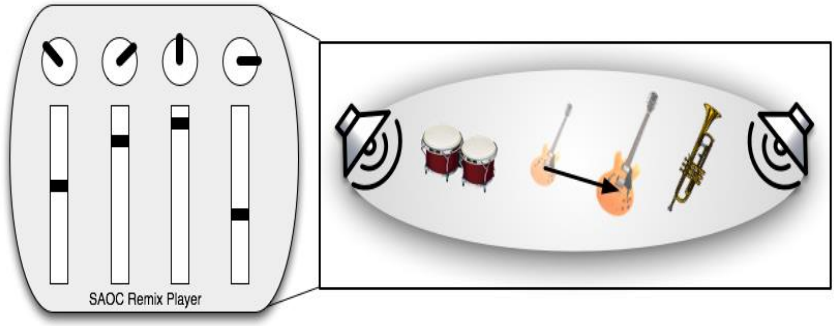
Tomado de: https://mpeg.chiariglione.org/standards/mpeg-d/mpeg-surround

MPEG – D Spatial Audio Object Coding (SAOC)

Diseñado para transmitir una serie de **objetos de audio**, en una señal conjunta

Crea una descripción paramétrica de las propiedades perceptivamente relevantes de los objetos de audio

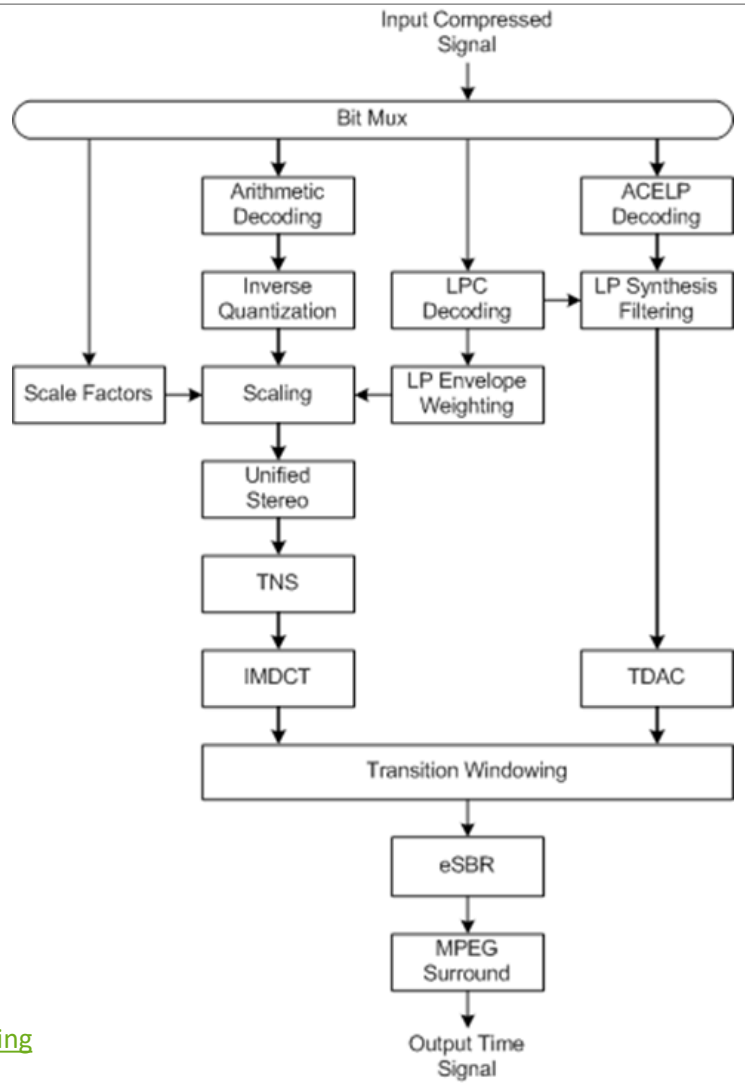




Tomado de: White Paper on MPEG Technology: Spatial Audio Object Coding; ISO/IEC JTC1/SC29/WG11 MPEG2015/N15820; October 2015, Geneva, Switzerland

MPEG-D Unified Speech and Audio Coding (USAC)

Diseñado para comprimir cualquier contenido compuesto de voz, música o una mezcla de voz y música.



Tomado de: https://mpeg.chiariglione.org/standards/mpeg-d/unified-speech-and-audio-coding

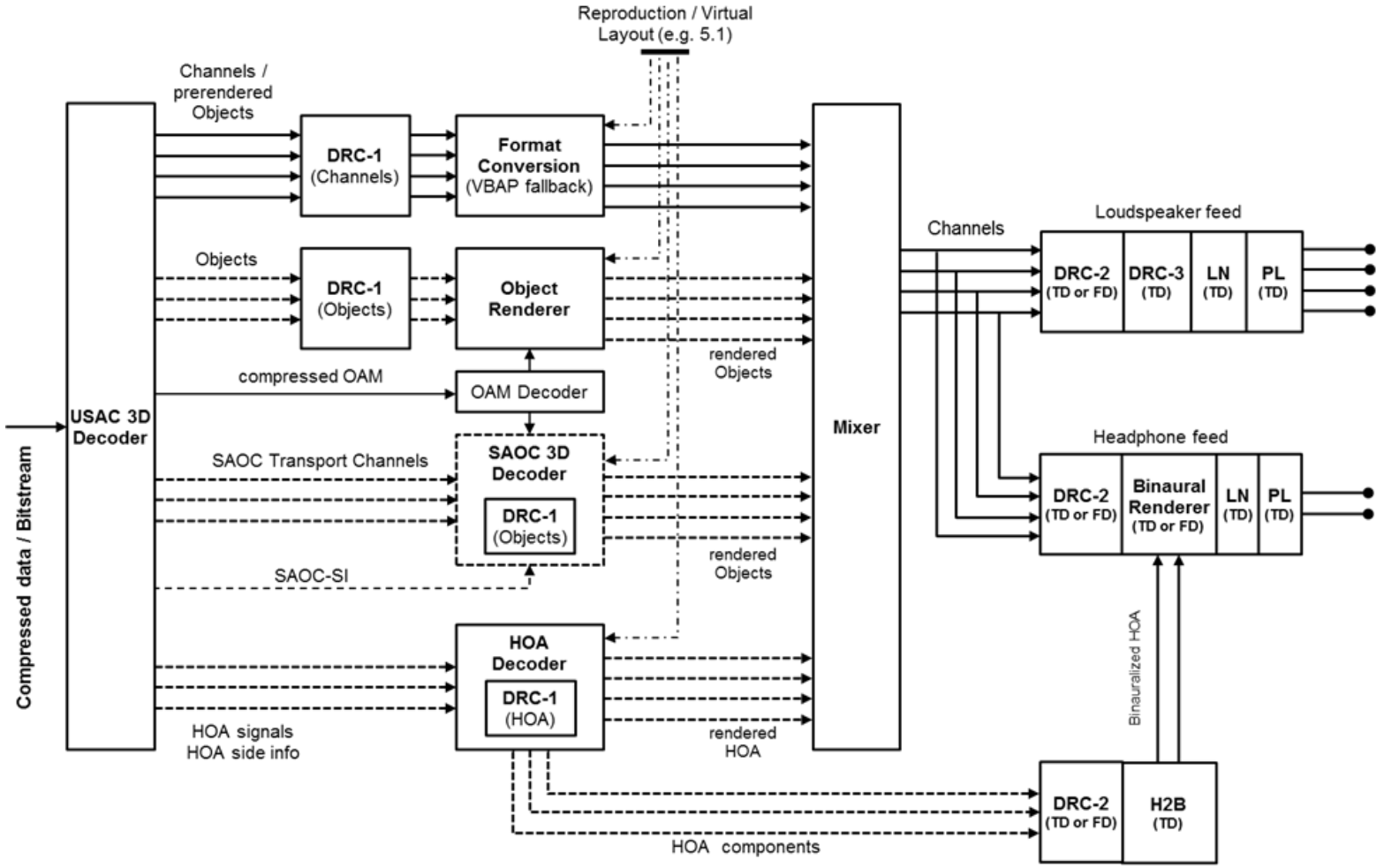
MPEG-H 3D Audio

Diseñado para escenarios en los que es necesario comprimir un programa de audio multicanal (por ejemplo, un programa de 22.2 canales) y presentarlo en los altavoces apropiados

- Ideal para Home Theaters

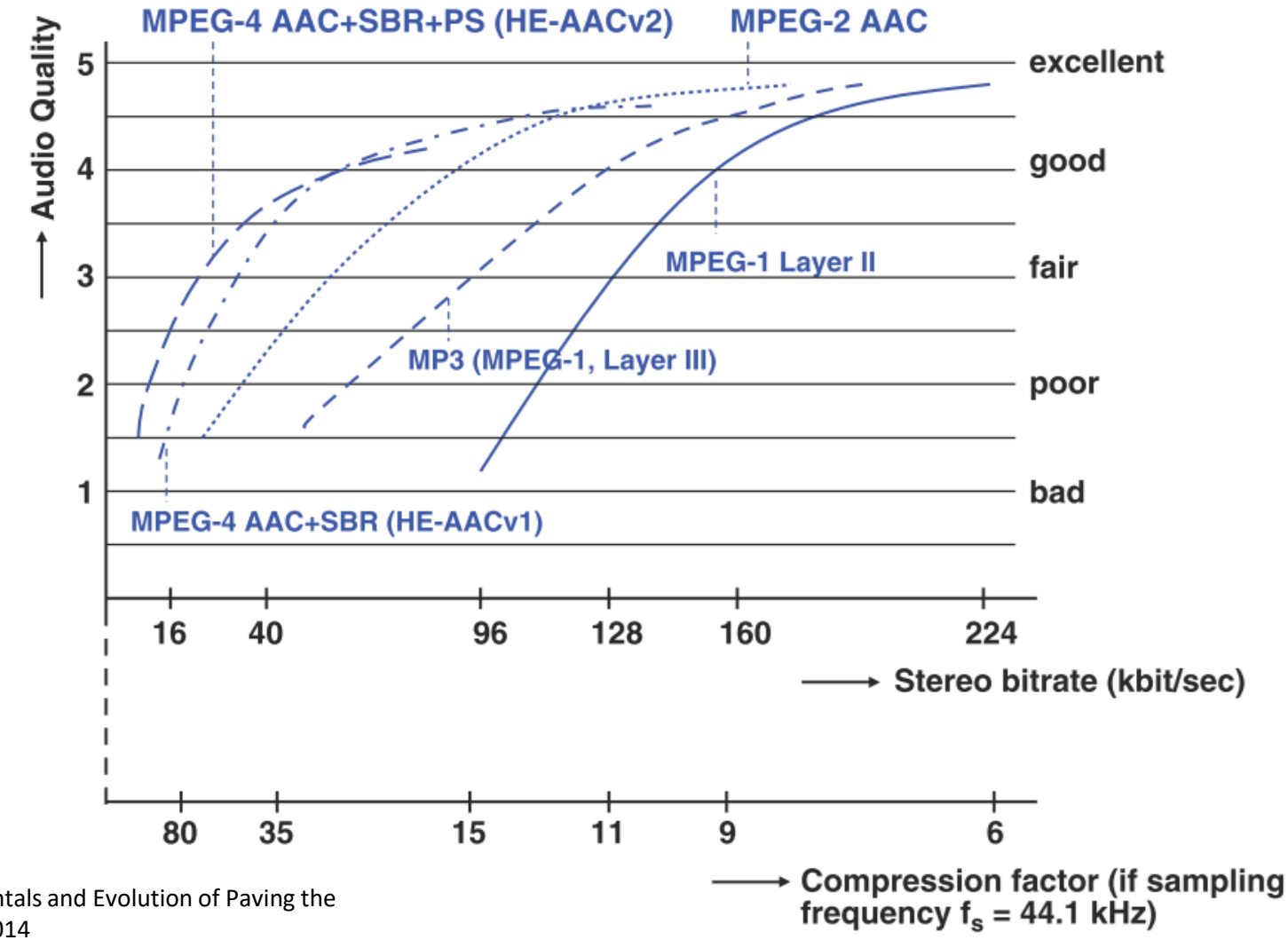
Tiene como objetivo mantener la sensación “envolvente” y de localización sonora precisa, incluso para sistemas que tienen pantallas pequeñas (tablets) con altavoces integrados en el dispositivo y/o con auriculares, y con canales de transmisión con bajas tasas de bits

MPEG-H 3D Audio



Tomado de: ISO/IEC CD 23008-3 - High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio

Comparación de codecs de audio



Tomado de MPEG-2 SYSTEMS Fundamentals and Evolution of Paving the MPEG Road, Jan van der Meer, Wiley, 2014

Videos recomendados

https://www.soundguys.com/mpeg-h-explained-24471/

https://www.youtube.com/watch?v=KOzLE1Osaew

(Barbería virtual)

Bibliografía y material de referencia

Introduction to Digital Audio Coding and Standards, Marina Bosi and Richard E. Goldberg, Kluwer Academic Publishers

Fundamentals and Evolution of MPEG-2 SYSTEMS. Paving the MPEG Road. Jan van der Meer. Wiley, 2014

Psychoacoustics Facts and Models, Hugo Fastl and Eberhard Zwicker, Springer, 2007

The Theory Behind Mp3, Rassol Raissi, December 2002

Full HD Voice is Nearly Here, Jeff Hecht, IEEE Sepctrum, June 2015

Spectral Band Replication, a novel approach in audio coding, Martin Dietz et al, Audio Engineering Society, 2002

ITU-T CODERS FOR WIDEBAND, SUPERWIDEBAND, AND FULLBAND SPEECH COMMUNICATION, IEEE Communications Magazine, October 2009

Presentación: “Digitalización y codificación de Audio”, Pablo Flores Guridi, Rafael Sotelo, IIE, 2014

New WID on EVS Codec Extension for Immersive Voice and Audio Services, 3GPP TSG SA Meeting #77, S4-170745, June 2017

Generative Speech Coding with Predictive Variance Regularization, KLEIJN, W. Bastiaan, et al, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021. p. 6478-6482.

Codificación de Voz y Audio

¡MUCHAS GRACIAS!