Formatos de ficheros de sonido

En general todos los formatos están diseñados para almacenar sonidos en el rango de los 20 Hz a 20000 Hz, que es básicamente el conjunto de frecuencias que puede percibir el ser humano. Algunos formatos están orientados a música y otros abarcan cualquier tipo de sonido, como la voz, ruido e incluso música.

Audio digital

El sonido puede digitalizarse desde un micrófono, sintetizador, DAT, etc.. Para digitalizar sonidos sólo se requiere una tarjeta de sonido que admita entrada para micrófono y audio.

Para digitalizar un sonido se toman m muestras del mismo cada segundo y se guardan digitalmente. La cantidad de muestras que se toman por segundo se denomina ***frecuencia de muestreo****, mientras que la cantidad de información tomada por muestra se conoce como* ***tamaño de la muestra (bits/muestra)****.*

Las tres frecuencias de audio más utilizadas son:

- *Calidad de radio: 11025 Hz*
- *Calidad telefónica: 20050 Hz*
- *Calidad de CD: 44100 Hz*

El tamaño de la muestra varía entre 8 y 16 bits. Mientras más grande sea la muestra, mejor se representará el sonido. El valor de cada muestra se redondea al entero más cercano (cuantificación). Ésta puede producir ruidos de fondo y distorsionar el sonido.

Se debe usar audio digital:

- *Cuando no se tenga control completo acerca del hardware de reproducción*
- *Cuando se disponga de suficientes recursos para manejar los archivos digitales*
- *Cuando se necesite reproducir diálogos.*

Formatos de fichero para Música

Aunque se puede almacenar una canción en un fichero WAVE hay dos razones por las que los formatos de música especializados son importantes:

- *El tamaño: Es más compacto almacenar una lista de notas que grabar la canción entera.*
- *Facilidad de modificación: Si se tienen las notas ejecutadas por cada instrumentos, se pueden modificar. Sin embargo, en un fichero wave es casi imposible modificar el sonido de un solo instrumento cuando está grabado el resultado final.*

Esto tiene su importancia, por ejemplo, en el caso de juegos por ordenador, puesto que están diseñados para jugar durante horas, el espacio de almacenamiento es importante. También es importante que la música de fondo cambie a medida que el juego progresa, lo que es más fácil de hacer a partir de notas.

Los formatos musicales basados en notas tienen dos desventajas:

- *Se está limitado en la elección de instrumentos*
- *Se supone que se ejecuta en la escala occidental de 12 tonos.*

Es importante tener en cuenta que la calidad final no depende del fichero, sino de la tarjeta de sonido que lo ejecute y por otro lado, no pueden reproducir diálogos.

Formato Wav

Wav es el formato nativo de Windows y es uno de los más utilizados. Su estructura está basada en el formato de intercambio de ficheros (IFF) desarrollada originalmente por ELECTRONIC ARTS para uso en el AMIGA. IFF también es la base para el formato AIFF de Apple.

A partir de éste, Microsoft definió un formato de fichero denominado Resource interchange file format (RIFF). Los ficheros RIFF están organizados en trozos (chunks) anidados. Dos variantes del RIFF son los ficheros WAV y los AVI.

Un fichero WAVE es un tipo particular de fichero RIFF y todo fichero riff comienza con los caracteres RIFF. A continuación, 4 bytes y el tipo de código

TROZO RIFF

DESPLAZAMIENTO	*LONGITUD (BYTES)*	*NOMBRE*	*CONTENIDO*
0	*4*	***ChunkID***	*'RIFF'*
4	*4*	***ChunkSize***	*Longitud de fichero – 8*
8	*4*	***Format***	*'WAVE'*

TROZO FORMATO

DESPLAZAMIENTO	*LONGITUD (BYTES)*	*NOMBRE*	*CONTENIDO*
12	*4*	***Subchunk1ID***	*'FMT'*
16	*4*	***Subchunk1Size***	*Longitud de los datos 'fmt' (16 bytes)*
20	*2*	***Audioformat***	*1 PCM* *2 Microsoft ADPCM* *6 ITU A-Law* *7 ITU μ-Law* *20 ITU ADPCM G.723* *49 GSM* *64 ITU ADPCM G.721* *80 MPEG*
22	*2*	***NumChannels***	*Canales*
24	*4*	***SampleRate***	*Muestras por segundo: ej. 44100*
28	*2*	***ByteRate***	*Canales *bits/muestra/8*
32	*4*	***BlockAlign***	*Fm* alineamiento de bloque*
34	*2*	***BitsPerSample***	*8 o 16*

TROZO DE DATOS

DESPLAZAMIENTO	*LONGITUD (BYTES)*	*NOMBRE*	*CONTENIDO*
36	*4*	***Suchunk2ID***	*'DATA'*
40	*4*	***Subchunk2Size***	*Longitud del bloque de datos*
44	*Indefinido*	***Data***	*Datos*

Las muestras de 8 bits se guardan como UNSIGNED y las de 16 bits como ENTEROS CON SIGNO EN COMPLEMENTO A 2, CUYO RANGO ES -32768 A 32767

MIDI

El formato MIDI surgió para dar respuesta al problema de conectar distintos instrumentos musicales electrónicos entre sí (Teclados, sintetizadores, etc…). Hay que tener en cuenta que el MIDI,en sus comienzo, no es un formato de fichero, sino un protocolo de comunicación entre instrumentos musicales (incluido el ordenador).

En 1988 se definió el Formato de fichero MIDI estándar por la asociación de fabricantes MIDI.

Ficheros MIDI estándar

Un fichero MIDI es una serie de trozos (chunks). Estos trozos tienen el mismo formato que los utilizados por AIFF, IFF y WAVE. Cada trozo tiene un nombre de cuatro caracteres y un código de longitud de 4 bytes y algunos datos. Al contrario que en otros formatos, los trozos MIDI no están anidados.

Identificando los ficheros MIDI

El trozo MThd, que aparece al comienzo de cada fichero MIDI, es la mejor manera de identificarlo.

Trozo de cabecera MIDI

El trozo de cabecera (MThd) contiene unos pocos datos sobre el fichero MIDI. Todos estos valores están guardados en formato MSB:

Bytes	*Descripción*
2	*Tipo de fichero*
2	*Número de pistas*
2	*Formato de tiempo*

La información musical de un fichero MIDI está organizada en varias pistas, cada una para un instrumento, por ejemplo. Hay algunos conceptos que tienen que estar claros:

- *Los ficheros MIDI tipo cero sólo contienen una pista.*
- *Los ficheros MIDI tipo uno contienen múltiples pistas que se pueden ejecutar simultáneamente.*
- *Los ficheros MIDI tipo dos contienen múltiples pistas, pero sin asumir ninguna relación entre ellas (son bastante poco usados).*

Pistas (Tracks) MIDI

Hay que distinguir entre pista (track) y canal (channel) MIDI. Aunque lo normal es que un fichero multipista ejecute cada pista en un canal no existe una relación biunívoca entre ambos.

El formato de una pista MIDI es simplemente una lista de sucesos (events) MIDI, cada uno precedido por un tiempo (delta time).

Enteros de longitud variable

Para ahorrar espacio, los ficheros MIDI utilizan enteros de longitud variable para almacenar tiempos y otros valores críticos. El formato almacena 7 bits en cada byte. El MSb indica si es el último byte (MSB=0) o si hay más bytes a continuación (MSB=1).

Tiempos de intervalo

Los sucesos MIDI ocurren en ciertos instantes de tiempo. Hay dos formas de marcar dicha información. Se puede almacenar el tiempo absoluto en el que ocurre dicho suceso o marcar el intervalo entre eventos. Los ficheros MIDI utilizan esta segunda aproximación. Cada suceso está precedido por un número que indica el número de ticks de reloj que lo separan del suceso previo

La duración precisa de cada tick depende del formato de tiempo especificado en la cabecera y puede cambiar mediante eventos especiales.

Sucesos MIDI

Un suceso MIDI es un conjunto de datos que especifica una acción musical, como pulsar una tecla o liberarla. El primer byte del conjunto es un byte de estado, que especifica el tipo de acción y, cuando es necesario, el canal. Los bytes de estado siempre tienen el bit alto a uno. Los bytes restantes son los bytes de datos, que nunca tienen el bit alto a uno. Esta distinción es importante.

Convenciones sobre numeración

Tradicionalmente, los canales MIDI se numeran de 1 a 16 y los instrumentos MIDI de 1 a 128. Sin embargo, el código numérico va de 0 a 15 y de 0 a 127. Por tanto hay que restar uno al convertir entre el código MIDI y el código almacenado en el fichero.

Temporización MIDI

El primer paso para ejecutar un fichero MIDI es traducir los tiempos almacenados en el fichero en algo más util. Los ficheros MIDI utilizan dos técnicas para especificar la duración de un tick:

- *Si el formato de tiempo en la cabecera es negativo, el fichero MIDI especifica el tick utilizando la convención SMPTE.*
- *Si el formato de tiempo es positivo, especifica el tiempo utilizando el tempo musical.*

Temporización SMPTE

La sociedad de Ingenieros de Televisión e Imágenes en movimiento (SMPTE) utiliza una técnica estándar para especificar tiempos precisamente. Esta técnica cuenta el paso de horas, minutos, segundos y marcos. Un marco es la duración de un marco de televisión y varía de 1/24 a 1/30 de segundo. Esta técnica es ampliamente utilizada por ingenieros de audio y vídeo y es una herramienta importante para sincronización precisa de fuentes de audio y vídeo.

- *24 frames por segundo son los utilizados por el equipo de imágenes en movimiento.*
- *25 frames por segundo son los utilizados por PAL y SECAM.*
- *30 frames por segundo son los utilizados por NTSC*

Temporización musical

Los músicos prefieren utilizar el tempo, normalmente en beats por minuto (bpm). Un beat normalmente corresponde a una negra (quarter). El tempo varía normalmente de 80 bpm a 200 bpm. El valor por defecto para MIDI es 120 bpm.

Uno de los aspectos más confusos de midi es el gran número de términos empleados:

- *Tick MIDI: Los valores de tiempo se especifican en ticks*
- *Clock MIDI: Un clock MIDI es 1/24 de una quarter.*
- *Quarter note MIDI: La duración de una nota quarter MIDI (en milisegundos) está especificada por un metaevento de tiempo. El fichero MIDI estándar realcione las quarter con los tick MIDI.*
- *Quarter note MUSICAL: Normalmente una quarter note MIDI corresponde a una quarter note MUSICAL (pero no siempre).*
- *Tempo: Se mide en beats por minuto.*
- *Metrónomo: El ritmo apropiado de metrónomo queda definido por un metaevento de tiempo.*

CSOUND

***Csound** es un paquete de software que permite la generación y edición de sonido con el objeto de componer música y sonido Implementa un lenguaje de programación de alto nivel con objeto dellevar a cabo dicha composición..*

Historia

En la década de los sesenta se encuentran los preliminares del programa Csound. Se basa en los programas Music1, Music2, Music3 y Music4, desarrollados durante 1964 y 1965. La siguiente versión se denominó MusicB. Para evitar el problema de tener que rescribir el código con cada nueva máquina, Music4B se desarrolló completamente en Fortran. Después de algún tiempo apareción el Music11, que puede considerarse como el verdadero predecesor de Csound.

En la creación de Csound han intervenido compositores interesados por la música y la informática.

Funcionamiento del Csound

El Csound define dos objetos relevantes para la composición. Por un lado, la orquesta, compuesta de instrumentos y por otro la partitura, que especifica el orden en que se va a reproducir cada sonido.

Como ejemplo, podemos ver la generación de un tono sinusoidal. Aunque Csound tiene ya una manera de entrada de código compacta, mediante el formato csd, la comunicación con Csound se realiza mediante dos archivos en formato texto. Uno para la orquesta (orc) y otro para la partitura (sco). Una posible orquesta sería:

```
instr 1;intrumento 1
       iamplitud = 15000
       ifrecuencia = p4
       itabla = 1
       a1 oscil iamplitud, ifrecuencia, itabla
       out a1
endin;final de instrumento 1
```

Una posible partitura es la que se muestra a continuación. En ésta se reproducen tres notas. La primera de 440 Hz durante cuatro segundos a partir del comienzo, la segunda de 880 Hz a partir del segundo 3 durante 2 segundos. La tercera de 1760 Hz, desde el segundo 5 durante un segundo.

```
;instr  comienzo duración p4 (en este caso: frecuencia)
 i1       0        4        440
 i1       3        2        880
 i1       5        1        1760
 e      ;fin de la partitura
```

Audio MPEG

Codificación perceptual

En los sistemas de compresión habituales, se comprime toda la información existente. Si disponemos un modelo psicoacústico del del sistema auditivo humano, el codificador identifica el contenido imperceptible de la señal (partes irrelevantes) y las elimina, codificando de forma eficiente aquellas partes que sí son perceptibles.

Es necesario conocer la diferencia entre frecuencia y pitch. La frecuencia es una medida objetiva, sin embargo el tono (pitch) no lo es. Puede ocurrir que dos sinusoides con la misma frecuencia pero distinta amplitud se perciban con distinto pitch. Otro fenómeno a considerar es el batido de frecuencias. Éste es un fenómeno que aparece cuando se escuchan simultáneamente dos frecuencias muy cercanas. Cuando la diferencia de frecuencia entre ambas tiene un valor que corresponde a una frecuencia audible, esta diferencia se puede llegar a percibir de forma nítida. Aunque es discutible, algunos aseguran oir un tono suma (la suma de las frecuencias de ambos tono). También puede aparecer un intertono, especialmente a frecuencias inferiores a 200 Hz, cuando existen dos tonos simultáneos de baja intensidad. Por ejemplo, dos tonos simultáneos de 65 y 98 Hz no se oirán como un intervalo perfecto, sino como un tono de 82 Hz. Por otra parte, si se escuchan dos tonos de frecuencias inferiores a 500 Hz, uno a continuación del otro, el oído es capaz de detectar diferencias de frecuencias de hasta 2 Hz.

El oído tiene una gran rango dinámico. Se define el umbral de dolor como la intensidad de sonido a partir del cual se empiezan a producir daños en el oído y el umbral de audición como la intensidad el sonido más pequeño que podamos escuchar. La diferencia de intensidad entre el umbral de audición y el de dolor es de 1.000.000.000.000 de veces. Debido a esta diferencia se utiliza una escala logarítmica entre 0 y 120 dB SPL (sound pressure level. Un efecto importante es que la sensibilidad del oído depende de la frecuencia:

- *La máxima sensibilidad se encuentra entre 1 y 5 kHz, con cierta insensibilidad a las altas y bajas frecuencias.*
- *Las curvas de igual sonoridad muestran el margen de frecuencias percibido con la misma sonoridad. La curva inferior corresponde al umbral de audición. Por ejemplo, un tono de 30 Hz apenas audible, tiene una intensidad 60 dB mayor que un tono de 4 kHz de un nivel equivalente al umbral de audición. La respuesta varía en función de la intensidad. Cuanto más intenso sea el sonido más plana es la curva.*

No toda la información presente en una señal de audio es percibida por el oído. Se utiliza la Entropía perceptual como medida de la información presente en una señal. Las señales con poca entropía pueden ser codificadas perceptualmente de forma muy eficiente (gran reducción binaria). Por esta razón, un codificador de audio debería diseñarse con una tasa binaria de datos variable:Baja si la información es pobre y Alta si hay más información. La salida debe ser variable porque aunque la frecuencia de muestreo sea constante, la entropía no lo es.

La eliminación de las componentes no percibidas se conoce como ***reducción de datos****. La señal original no se puede reconstruir de forma exacta. La función de cualquier sistema de reducción de datos es eliminar (reducirla) la entropía de la señal.*

Fundamentos de la codificación perceptual

El objetivo de un sistema de reducción de datos es disminuir la tasa binaria: Ésta consiste en ***el producto de la frecuencia de muestreo por el tamaño de cada muestra.*** *Esta reducción se puede conseguir de varias maneras. Por un lado reduciendo la frecuencia de muestreo. Sin embargo, debido al teorema del muestreo, esto limita el rango de frecuencias de la señal de audio.Otra posibilidad es reducir el tamaño de cada muestra. Sin embargo, esto repercute en el margen dinámico de la señal. Se perderán 6 dB por cada bit (se incrementa el ruido de cuantificación). Por último, con técnicas psicoacústicas es posible reducir la tasa binaria sin perder margen dinámico ni contenido espectral de la señal.*

Estándar de audio MPEG-1

La organización Internacional de Estandarización (ISO) y la Comisión Internacional Electrónica (IEC) crearon en 1988 el ***Moving Pictures Expert Group*** *(MPEG) con el fin de que se formalizaran las*

*técnicas de compresión de audio y vídeo. Este grupo de expertos ha desarrollado varios estándares de enorme importancia. El primero de ellos fue el ISO/IEC 11172, "**Codificación de imágenes en movimiento y audio asociado para la grabación digital hasta 1.5 Mbits/s**". Se finalizó en noviembre de 1992. A este estándar se le conoce normalmente como MPEG-1. El estándar está formado por cuatro partes:*

1. *El sistema (audio y vídeo)*
2. *El vídeo*
3. *El audio. El máximo régimen binario para audio es de 1.856 Mbps.*
4. *pruebas de conformidad*

La parte del estándar dedicada al audio (ISO/IEC 11172-3) ha encontrado multitud de aplicaciones, como el CD-RO, por ejemplo. Permite codificaciones de datos PCM con frecuencias de muestreo de 32, 44.1 y 44.8 kHz, a tasas binarias comprendidas entre 32 y 224 kbps/canal (entre 64 y 448 kbps para canales estéreo). Puesto que las redes digitales de datos utilizan tasas binarias de 64 kbps (8 bits a 8 kHz), la mayoría de los codificadores presentan tasas binarias múltiplos de 64.

El estándar ISO/MPEG-1 fue desarrollado específicamente para codificar audio y vídeo sobre el formato CD, con el régimen binario que soporta (1.41 Mbps). Sin embargo, el estándar soporta tasas binarias para señales estéreo entre 64 kbps y 448 kbps, incluyendo 32 kbps en señales mono. El estándar MPEG-1 está basado en algoritmos de compresión de datos cuya investigación y desarrollo se remonta a varias décadas atrás.

*El sistema MUSICAM (Masking pattern Universal Subband Integrated Coding and Multiplexing) fue uno de los primeros y más eficientes algoritmos de codificación perceptual. Basado en el sistema MASCAM (Masking pattern adapted Subband Coding and Multiplexing) el sistema MUSICAM divide la señal en 32 subbandas y utiliza modelos de codificación perceptual, con **umbrales de audición y enmascaramiento**, para conseguir una determinada tasa de reducción binaria. Con una frecuencia de muestreo de 48 kHz, cada subbanda tiene un ancho de 750 Hz, un **factor de escala** de 6 bits, cuyo valor depende del máximo de cada grupo de 12 muestras. Éstas, a su vez, son cuantificadas con una longitud binaria comprendida entre 0 y 15 bits. Los factores de escala se calculan sobre un intervalo de 24 ms, el correspondiente a un grupo de 36 muestras. Sólo se codifican aquellas subbandas que contengan señales audibles, por encima del **umbral de enmascaramiento**. Las subbandas con señales que tengan un nivel muy superior al umbral de enmascaramiento se codifican con un mayor número de bits, aumentando su relación S/N. Además, de forma paralela, se realiza un análisis espectral con la transformada de Fourier que permita determinar los umbrales de enmascaramiento. De esta forma, la tasa binaria se reduce a un valor de 128 kbps por canal. Numerosas pruebas realizadas han demostrado:*

1. *Que el sistema MUSICAM consigue una calidad equiparable al CD*
2. *Que es compatible con señales mono y estéreo.*
3. *Que no existe degradación cuando se utilizan dos codificadores en cascada.*
4. *Que su calidad es muy superior a la de las señales FM*

*La parte del estándar ISO/MPEG-1 correspondiente al audio fue probada por primera vez en julio de 1990 por la radio nacional sueca. El sistema MUSICAM fue considerado superior en términos de complejidad de diseño y por el bajo retardo de codificación introducido. Sin embargo, los codificadores por transformada del tipo ASPEC (**Adaptive Spectral Perceptual Entropy Coding**) proporcionan una mejor calidad a tasas más bajas. Las arquitecturas de estos dos tipos de codificación están en la base del estándar ISO/MPEG-1. El estándar 11172-3 describe tres capas de codificación, conocidas como **layers**, cada una de ellas destinadas a diferentes aplicaciones.*

1. *El layer I describe la codificación más simple con unas tasas binarias relativamente altas (192 kbps/canal)*
2. *El layer II está basado en el layer I, aunque es algo más complejo, y trabaja a unos régimenes inferiores (96-128 kbps) El layer II-A es una versión **joint-stereo** con regímenes binarios comprendidos entre 128 y 192 kbps por cada pareja de canales estéreo.*
3. *El layer III es conceptualmente diferente de los layer I y II. Es el más sofisticado y con menor tasa binaria (64 kbps/canal).*

Modelos psicoacústicos

El estándar MPEG-1 propone dos modelos psicoacústicos para determinar los umbrales de enmascaramiento mínimos que aseguran la audibilidad de la señal. Por ejemplo, el modelo 1 realiza los siguientes pasos:

1. *Conversión tiempo-frecuencia: Se utiliza una FFT de 512 o 1024 puntos con una ventana de Hanning para reducir los efectos de borde y transformar los datos al dominio de frecuencia.*
2. *Determinación de los niveles máximos SPL. Este cálculo se realiza en cada subbanda utilizando los factores de escala y los coeficientes espectrales de la FFT. Los valores máximos son considerados como potenciales señales de enmascaramiento y servirán para determinar posteriormente los umbrale de enmascaramiento.*
3. *Determinación de los umbrales de audición en ausencia de señal. El umbral absoluto de audición se determina en ausencia de cualquier señal de audio. Este umbral será el mínimo umbral de enmascaramiento.*
4. *Identificación de componentes tonales y no tonales. Se identifican las componentes tonales (sinusoidales) y no tonales (ruido) y se procesan de forma independiente, ya que las curvas de enmascaramiento son distintas en ambos casos.*
5. *Diezmado de las señales de enmascaramiento. Se reduce el número de componentes enmascaradotas, dejando sólo las más relevantes, en función de su amplitud y su separación espectral en* ***barks****.*
6. *Cálculo de los umbrales de enmascaramiento. Se determinan en cada subbanda los umbrales de enmascaramiento, producidos por las componentes enmascaradotas ruidosas, aplicando una determinada curva de enmascaramiento*
7. *Determinación de la curva de enmascaramiento global. Es la suma de todas las curvas de enmascaramiento que han sido asignadas a cada subbanda. La curva final también tiene en cuenta el umbral de audición en ausencia de señal.*
8. *Determinación de los umbrales mínimos de enmascaramiento en cada subbanda.*
9. *Cálculo de la relación señal a enmascaramiento (SMR) para cada subbanda (diferencia entre el máximo SPL y los umbrales mínimos de enmascaramiento).*

Layer I

El layer I es una versión simplificada del estándar MUSICAM. El objetivo de este layer es proporcionar una alta calidad a un bajo coste, aunque con una elevada tasa binaria.

- *La señal de entrada se divide en 32 subbandas de igual ancho con un banco de filtros polifase. El filtro es muestreado de forma crítica; el número de muestas a la salida del banco de filtros es igual al número de muestras de entrada. Las subbandas contiguas se encuentran solapadas en frecuencia. Cada uno de los filtros de los correspondientes filtros inversos son sistemas sin pérdidas (reconstrucción perfecta de la señal). Sin embargo, la cuantificación de datos produce un pequeño error en la reconstrucción de la señal. Todas las bandas del banco de filtros son iguales, pero las* ***bandas críticas de nuestro oído*** *no lo son. El algoritmo de asignación binaria tiene que tener en cuenta este detalle y realizar la oportuna compensación. Por ejemplo, a las bandas más bajas se les suele asignar un mayor número de bits.*
- *El banco de filtros entrega 32 muestras (una por canal) por cada 32 muestras de entrada. En el layer I se forma tramas tomando como base 12 muestras de cada una de las 32 bandas. Esto representa un total de 384 muestras. En cada grupo de 12 muestras se realiza una determinada* ***asignación de bits****. Las subbandas que hayan sido consideradas inaudibles tienen una asignación nula. Teniendo en cuenta el umbral de enmascaramiento, el algoritmo de asignación determina el número de bits necesario para cuantificar las muestras. Para codificar éstas, se utiliza una notación en coma flotante.*
- *Cada muestra se codifica con un código PCM. El cuantificador proporciona* 2^{n-1} *intervalos de cuantificación (2<=n<=15). Las subbandas con altas SNR son las que*

tienen las muestras con mayor longitud binaria y las subbandas con bajas SNR tienen muestras con menor longitud binaria.

Layer II

El layer II es esencialmente idéntico al MUSICAM, y por tanto similar al layer I, pero su diseño es más sofisticado. Intenta dar una mayor calidad de audio con regímenes binarios moderados, con un coste algo superior.

- *El banco de filtros crea 32 subbandas de igual ancho, pero el tamaño de la trama se triplica a 3x12x32 (1152 muestras por canal de audio). El análisis espectral se hace más preciso con una FFT de 1024 puntos. Los niveles de enmascaramiento se estiman teniendo en cuenta las componentes tonales y las no tonales.*
- *Por cada grupo de 12 muestras se realiza una única **asignación binaria**. En cada subbanda se calculan tres factores de escala, uno por cada grupo de 12 muestras. La asignación binaria cubre un gran abanico de niveles, desde 3 a 65535 (o ninguno), pero el número de niveles depende de cada subbanda. Las subbandas más bajas pueden llegar a tener 15 bits, las medias hasta 7 bits y las altas están limitadas a 3 bits.*
- *En cada banda, las señales más fuertes son las que tienen mayores longitudes binarias. La cuantificación, así pues, varía con el orden de la subbanda. Las subbandas más altas normalmente reciben menos bits y por tanto, los pasos de cuantificación son mayores. Además, con objeto de tener una codificación más eficiente, las muestras se agrupan de forma sucesiva de tres en tres (en cada una de las 32 subbandas) formando los denominados **gránulos**.*

Layer III

El layer III combina elementos precedentes de MUSICAM y ASPEC, y su diseño es mucho más complejo que los layers I y II. La calidad de la señal de audio es moderada, incluso con regímenes binarios bastantes bajos. Los ficheros de audio codificados con el layer III son también conocidos como MP3.

- *Al igual que en los layers I y II, la señal de audio se divide en 32 subbandas por un banco de filtros polifase. Además, las muestras de cada subbanda se transforman en 18 coeficientes espectrales utilizando una MDCT (transformada discreta coseno modificada) dando un máximo de 576 coeficientes con una resolución espectral de 41,67 Hz a una frecuencia de muestreo de 48 kHz. La resolución temporal es de 24 ms. La resolución espectral es óptima en señales estacionarias, aunque a expensas de una resolución temporal, necesaria cuando aparecen transitorios de nivel en la señal. Los coeficientes espectrales se agrupan en bandas de un ancho similar al de las bandas críticas, cada una de ellas con un determinado factor de escala. Con frecuencias de muestreo inferiores, como las de MPEG-2, la resolución espectral se puede llegar a duplicar e incluso más. Por ejemplo con 24 kHz de frecuencia de muestreo, la resolución espectral es de 21 Hz. Esta resolución tan baja permite una mejor adaptación de los factores a la anchura real de las bandas críticas De esta forma se consigue una buena calidad de audio con bajas tasas binarias, a expensas de mayor resolución espectral.*
- *Los errores de cuantificación pueden aparecer en forma de precoz (chasquidos). Por esta razón, la longitud de las ventanas MDCT se cambia dinámicamente para tener una mejor resolución espectral o temporal, según el caso. La longitud de las ventanas más largas corresponde a 36 muestras y se utiliza cuando las componentes de la señal son estacionarias. Las ventanas más cortas son de 12 muestras y se utilizan cuando la señal presenta transitorios de alto nivel. La resolución temporal en este caso es de 8 ms a una frecuencia de muestreo de 48 kHz y el número de coeficientes espectrales es de 192.*
- *El algoritmo de asignación binaria realiza una cuantificación dinámica de las muestras de audio: Un bucle iterativo de asignación calcula el ruido de cuantificación en cada una de las subbandas. A este proceso se le llama asignación de ruido, en contraste con el proceso de asignación de bits. Mediante una técnica de análisis síntesis se estima si el espectro*

cuantificado cumple los requerimientos de la curva de enmascaramiento. La cuantificación es no uniforme. Los coeficientes se amplifican antes de cuantificarlos, optimizando la SNR. A continuación se aplica un codificación Huffman, tanto para los coeficientes como para los factores de escala, para eliminar las redundancias estadísticas y conseguir una mayor compresión binaria.

- *En el layer III la tasa binaria puede variar de una trama a otra, es decir es una grabación de tasa variable. De esta forma, a los pasajes con menos música se les puede asignar menos bits y donde haya más información se le asignan más bits. Cuando se trabaja con una tasa binaria fija, el sistema de reserva de bits codifica de forma óptima las partes de la señal que requieran una gran asignación de bits. La tasa binaria nunca puede exceder la capacidad del canal.*
- *El layer III trabaja opcionamente con los modos de codificación estéro MS (mid/side) y estéreo intenso. En una grabación estéreo intensa los canales L y R no se codifican de forma independiente. Realmente se transmite una sola señal (L o R) y la dirección de la imagen estéreo. El modo de codificación puede variar de trama a trama.*