INTRODUCCIÓN A LA CODIFICACIÓN DE AUDIO NATURAL A BAJAS TASAS BINARIAS. DE MPEG-1 A MPEG-4

Enrique Alexandre

E.T.S.E. de Telecomunicación
Universidad de Vigo
ealex@gts.tsc.uvigo.es

Antonio S. Pena

E.T.S.E. de Telecomunicación
Universidad de Vigo
apena@tsc.uvigo.es

RESUMEN

En este artículo se pretende ofrecer una visión general del estado actual del arte en la codificación de audio, introduciendo algunas técnicas básicas empleadas en la gran mayoría de los esquemas existentes. Nos centraremos especialmente en los estándares que el grupo ISO/IEC MPEG ha venido desarrollando desde 1988 y que pueden considerarse como la principal referencia a nivel mundial por su amplio margen de aceptación (Un buen ejemplo lo representa el tan extendido formato de ficheros mp3, formalmente conocido como MPEG-1/2 Capa III).

1. INTRODUCCIÓN

Se intentarán introducir ahora algunos conceptos muy básicos relacionados con la codificación de señales de audio, sin entrar en detalles de implementación ni a discutir las complejidades de los esquemas, por salirse estos temas fuera de los objetivos de este tutorial.

1.1. Redundancia Vs. Irrelevancia

A la hora de reducir la información necesaria para representar una determinada señal se puede optar por eliminar la redundancia presente en la misma o bien la irrelevancia. El primer caso sería el equivalente, por ejemplo, a representar un tono puro por sus valores de amplitud, frecuencia y fase en vez de por todas las muestras PCM que lo constituyen. Esta es la alternativa que utilizan los codificadores sin pérdidas, en los que la señal reconstruida es exactamente igual a la señal de entrada, sin ningún tipo de distorsión [1][2]. Sin embargo presentan el problema de requerir una elevada carga computacional para conseguir resultados aceptables.

Pero se puede tomar un segundo camino (no excluyente del anterior), y es aprovechar el hecho de que la señal de audio que pretendemos codificar va a ser escuchada por personas, y que conocemos o por lo menos podemos modelar hasta cierto punto la forma en que la mayoría de las personas oyen [3][4][5]. Se trata pues, de modelar el sistema auditivo humano (HAS, *Human Auditory System*), y a partir de este modelo caracterizar qué partes de la señal a codificar resultan más relevantes que otras y deben o no deben ser codificadas para mantener una calidad subjetiva dada. La señal reconstruida en este caso *no* será igual a la señal original pero, si el proceso se realiza correctamente, perceptualmente no se notará ninguna diferencia entre ambas. Estos codificadores pueden reducir todavía más la cantidad de información necesaria para representar la señal, hasta menos de 1 bit por muestra frente a los 10 bits necesarios en general por un codificador sin pérdidas [6].

1.2. El fenómeno del enmascaramiento

El fenómeno principal que modelan casi todos los sistemas existentes es el conocido como enmascaramiento simultáneo. Podríamos entenderlo como sigue. Imaginemos que estamos en un paraje solitario, donde sólo se oye el trino de los pájaros. Ahora imaginemos que aparece una cascada en este mismo sitio. Es fácil imaginarse que dejaremos de oír a los pájaros y sólo oiremos el ruido del agua, pero eso no significa que los pájaros no sigan ahí, sino tan sólo que su ruido ha quedado *enmascarado* por el de la cascada.

Otro tipo de enmascaramiento que también se da es el temporal, que consiste en que varios milisegundos antes y varias decenas de milisegundos después de un ruido se va a producir un cierto enmascaramiento, que, aunque es más difícil de modelar que el simultáneo, es de vital importancia para el caso de tratar con señales con un perfil temporal muy abrupto (ataques de castañuelas, golpes, etc.).

Se han realizado gran cantidad de estudios sobre estos fenómenos y otros más complejos [7][8][9], que han llevado a la existencia de distintos modelos psicoacústicos que serán los que utilice el codificador de audio para decidir qué partes de la señal resultan más importantes subjetivamente.

1.3. Esquema básico de un codificador

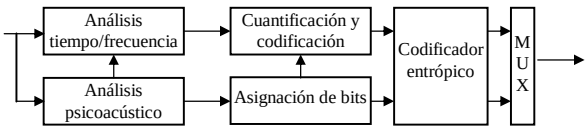

Figura 1. *Esquema general de un codificador perceptual de audio.*

En la Figura 1 se muestra un esquema general de un codificador perceptual de audio. Se pueden distinguir los siguientes bloques principales:

- Banco de filtros: Se utiliza para descomponer la señal de entrada en componentes espectrales submuestreadas. Junto con el banco de filtros correspondiente en el decodificador forma un conjunto de análisis/síntesis.
- Cuantificación y codificación: Este es el verdadero *núcleo* del sistema, donde las componentes espectrales se cuantifican y codifican intentando mantener el ruido siempre por debajo del umbral de enmascaramiento, tal y como se comentó anteriormente.
- Modelo perceptual: Proporciona toda la información necesaria para la correcta codificación de la señal manteniendo un determinado nivel de calidad requerido.
- Entramado: El último paso será siempre generar una trama binaria en la que se incluye toda la información y cabeceras necesarias para que el codificador sea capaz de interpretarla correctamente.

1.4. Esquemas fijados o parcialmente fijados

Dentro de los diversos estándares existentes en relación a la codificación de audio y atendiendo a la forma en la que se especifica el diseño a seguir para construir el codificador, podemos distinguir dos filosofías distintas. En primer lugar estarían aquellos esquemas, como el ITU-T G.722, en los cuales la estructura del codificador se encuentra perfectamente definida en la norma, no permitiendo ningún tipo de cambio sobre la misma.

Otra aproximación sería la seguida por ejemplo en los estándares MPEG, que consiste en fijar ciertos bloques del codificador necesarios para mantener la compatibilidad, dejando abierta la estructura del resto de los componentes del sistema. Esto hace que dos codificadores MPEG distintos puedan producir resultados totalmente opuestos en las mismas condiciones siendo ambos totalmente normativos.

2. ESTÁNDARES MPEG DE CODIFICACIÓN DE AUDIO

El grupo MPEG (ISO/IEC JTC1/SC29/WG11 MPEG, *Moving Pictures Expert Group*) ha desarrollado distintos estándares para la codificación de imágenes en movimiento, su información de audio asociada y su combinación desde 1988 [11-14]. Se pueden citar múltiples aplicaciones de los estándares de codificación de audio MPEG, tales como:

- DAB (*Digital Audio Broadcasting*).
- Transmisión RDSI.
- Audio de acompañamiento para la TV digital.
- *Internet streaming*.
- Audio portátil.
- Almacenamiento e intercambio de ficheros de audio en ordenadores.

Los formatos más utilizados actualmente son MPEG-1/2 Audio Capas II y III. Además, muchos sistemas en desarrollo prevén usar MPEG-2 AAC como sistema de codificación de audio.

2.1. MPEG-1

El proceso de desarrollo de MPEG-1 duró 4 años, desde 1988 hasta 1992, para finalmente convertirse en la norma ISO/IEC 11172. La parte del estándar relativa a la codificación de audio, ISO/IEC 11172-3 [10], describe un sistema adaptable a distintas aplicaciones. Así, se describen tres *capas* distintas, cada una de ellas con un grado de complejidad distinto, siendo la más compleja la capa III (el conocido sistema de archivos mp3), la cual está optimizada para proporcionar la máxima calidad a tasas binarias en torno a 128 kbit/s para una señal estéreo.

2.2. MPEG-2 BC

En 1994 se finalizó el estándar MPEG-2 BC (*Backwards Compatible*) [11], que no hace más que modificar ligeramente algunas características de MPEG-1. Por una parte proporciona la posibilidad de utilizar frecuencias de muestreo más bajas (16, 22.05 y 24 KHz), con lo cual se pueden codificar señales a tasas binarias mucho más bajas aun a costa de reducir su ancho de banda. Por otra parte, se introduce la capacidad, completamente compatible hacia atrás, de transmitir señales multicanal, incluyendo el formato 5.1.

2.3. MPEG-2 AAC

Poco después de 1994 se pudo observar, en los tests realizados, que para incrementar de manera sustancial la eficiencia de la codificación era necesario plantear un esquema que no estuviese lastrado por la necesidad de ser compatible con los estándares anteriores. Por esto se creó un nuevo esquema, denominado MPEG-2 NBC (*Non-Backwards Compatible*), también conocido como AAC (*Advanced Audio Coding*) [12], el cual es capaz de proporcionar una misma calidad subjetiva que los esquemas anteriores con un consumo de bits notablemente inferior. No obstante se produce un claro aumento de la complejidad computacional, incluso en el perfil más básico descrito en la norma.

2.4. MPEG-4

La primera versión de MPEG-4 se finalizó en 1998 [13], y una segunda versión se editó en 1999 [14]. MPEG-4 ya no proporciona un esquema de codificación en sí, sino que centra sus miras en conseguir proporcionar nuevas funcionalidades. El siguiente apartado estará dedicado íntegramente a este estándar.

2.5. MPEG-7

Al margen de la codificación de audio, pero relacionado con ella, está el estándar MPEG-7, pendiente de aprobarse en Julio de 2001, que pretende fijar las pautas para la búsqueda, filtrado, manejo y procesado de información multimedia.

3. MPEG-4. CODIFICACIÓN DE AUDIO NATURAL

El estándar ISO/IEC MPEG-4, en su parte 3 (Audio), pretende proporcionar un conjunto de tecnologías que permitan satisfacer las necesidades de autores, proveedores de servicio y usuarios

finales al mismo tiempo. Se facilita un completo conjunto de aplicaciones que van desde la codificación de voz hasta codificación de audio multicanal de alta calidad, tanto para sonidos naturales como sintetizados [15][18]. En particular, se permite la representación eficiente de objetos de audio consistentes en:

- *Señales de voz*: Se puede codificar con tasas binarias desde 2 kbit/s hasta 24 kbit/s. También se pueden conseguir tasas binarias más bajas en media utilizando herramientas de tasa binaria variable. Se permiten esquemas de bajo retardo para comunicaciones y, si se utilizan las herramientas HVXC, la velocidad y el *pitch* pueden ser modificados en tiempo real por el usuario.
- *Voz sintética*: Existen codificadores TTS /*Text-To-Speech*) con velocidades binarias entre 200 bit/s hasta 1.2 kbit/s, que admiten como entrada texto, o texto con información prosódica para generar voz sintética inteligible. Hay que notar que en este caso el estándar se limita a definir un formato de intercambio, sin entrar en las propias técnicas de conversión texto-voz.
- *Señales de audio natural*: Permite la codificación de señales de audio natural desde velocidades binarias muy bajas hasta alta calidad utilizando técnicas de transformada. Empieza con 6 kbit/s con 4 KHz de ancho de banda, pero permite calidad de difusión para mono o multicanal. Para la codificación de audio de alta calidad se ha adoptado el esquema MPEG-2 AAC con ligeras modificaciones y alguna que otra herramienta nueva.
- *Audio sintético*: El audio sintético viene proporcionado por el denominado *Decodificador de Audio Estructurado*.
- *Audio sintético de complejidad limitada*.

Uno de los puntos en los que más énfasis se hace en la norma de MPEG-4 es el de la escalabilidad. MPEG-4 proporciona cuatro tipos de escalabilidad, lo que hace que sea un esquema extremadamente flexible y adaptable a múltiples aplicaciones, así como, si pensamos en la transmisión sobre IP, se puede adaptar a las diferentes incidencias del tráfico. Los cuatro tipos de escalabilidad serán los siguientes:

- *Bitrate*: Una trama binaria se puede convertir a una trama binaria de tasa binaria inferior y todavía ser decodificada sin ningún problema por medio del uso de diferentes capas dentro de la estructura binaria. Este proceso se puede realizar tanto durante la transmisión como en el propio decodificador.
- *Ancho de banda*: Es un caso particular de escalabilidad del *bitrate*, donde una parte de la trama binaria que se corresponde con una banda frecuencial puede ser descartada durante la transmisión o la decodificación.
- *Complejidad del decodificador*: Permite que haya codificadores de diferentes complejidades todos ellos generando tramas binarias perfectamente válidas.
- *Complejidad del decodificador*: Permite que una misma trama binaria sea interpretada por decodificadores con distintos grados de complejidad. En general, la calidad final de la señal dependerá de la complejidad tanto del codificador como del decodificador.

Otras características no menos importantes incluidas en MPEG-4 serían las siguientes:

- *Modificación de la velocidad*: Permite variar la escala temporal sin alterar el *pitch* de la señal durante la decodificación. La aplicación más inmediata sería proporcionar un botón de avance rápido de la señal, o adaptar la señal de audio a una determinada secuencia de video.
- *Modificación del pitch*: Permite alterar el *pitch* sin variar la escala temporal durante la decodificación. Se puede utilizar para alterar la voz o para aplicaciones de karaoke. Esta técnica solo se utiliza en codificación paramétrica o audio estructurado.
- *Efectos de audio*: Proporcionan la posibilidad de procesar una señal decodificada con total precisión temporal para mezclarla, reverberarla, especializarla, etc.

4. MPEG-4 Versión 2

La primera versión de MPEG-4 fue finalizada en Octubre de 1998, pero el trabajo en este estándar continuó, cristalizando en una segunda versión en Diciembre de 1999. Todas las herramientas y perfiles existentes en la versión 1 no se reemplazan en la versión 2, sino que se han añadido a MPEG-4 nuevos perfiles, manteniéndose la compatibilidad hacia atrás con la versión 1. En la Figura 2 se esquematiza la relación entre las dos versiones [18].

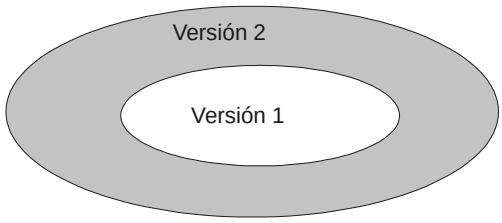

Figura 2. *Relación entre las versiones de MPEG-4.*

En MPEG-4 Audio Versión 2 se añaden las siguientes funcionalidades:

- *Robustez frente a los errores*: Se proporciona robustez frente a los errores en canales de transmisión. Los algoritmos de MPEG-4 Versión 2 clasifican cada campo de la trama binaria según su sensibilidad frente a los errores, de modo que la trama se divide en distintas clases que pueden ser tratadas por separado por las herramientas de protección de errores.
- *Codificación de bajo retardo*: El retardo del codificador MPEG-2/4 AAC puede llegar a ser de varias centenas de milisegundos, lo cual resulta del todo inaceptable para mantener una comunicación bidireccional en tiempo real. El nuevo perfil añadido a este codificador permite reducir el retardo hasta menos

de 20 ms, por medio de una reducción en el tamaño de la trama y la limitación en el uso de algunas herramientas.

- *Escalabilidad fina*: La escalabilidad del bitrate comentada antes hacía uso típicamente de una capa básica de 24 kbit/s junto con dos capas de mejora de 16 kbit/s cada una. En la versión 2 se permite una escalabilidad más fina, en pasos de tan sólo 1 kbit/s para un canal mono.
- *Codificación paramétrica de audio*: Las técnicas de codificación paramétrica están concebidas para conseguir codificar señales de audio a tasas binarias muy bajas, a partir de 4 kbit/s. En MPEG-4 se utiliza la técnica HILN (*Harmonic and Individual Lines plus Noise*) que consiste básicamente en descomponer la señal de audio en tonos aislados, patrones armónicos y componentes ruidosas, objetos que pueden ser representados de forma paramétrica con muy poca información. Este esquema se puede combinar en MPEG-4 con el utilizado para voz (HVXC), de modo que utilizando una herramienta de clasificación voz/música es posible seleccionar automáticamente HVXC para voz y HILN para música.
- *Espacialización ambiental*: Permite la composición de una escena de audio modelando de forma más natural que en la versión 1 la fuente sonora y el entorno acústico.
- *Canal de retorno*: Este canal permite peticiones de un cliente a un servidor, haciendo posible un servicio interactivo.
- *Trama de transporte de audio*: Se trata de un entramado especial que evita tener que recurrir al entramado general de MPEG-4 cuando la señal a transmitir es de audio.
- *Compresión de silencios CELP*: Se utiliza para reducir la tasa binaria gracias a una compresión eficiente de los silencios, que deben ser convenientemente detectados

5. CONCLUSIONES

Se ha pretendido dar un repaso general al estado del arte en el campo de la codificación de audio, centrándonos en el que ha venido a ser y sigue siendo el referente a nivel mundial, como es el conjunto de estándares MPEG. Dentro de éstos nos hemos encontrado con una marcada evolución de los codificadores de audio hacia estructuras cada vez más potentes a la vez que complejas, y que permiten obtener calidades iguales a velocidades cada vez más reducidas. El estándar MPEG-4 ha centrado la segunda parte del artículo, por representar el presente (y futuro) y aportar una visión conjunta e integradora de un gran número de técnicas de codificación, no solo de audio natural, como hemos podido ver.

6. REFERENCIAS

[1] Gerzon et al. *The MLP lossless compression system*. AES 17th International Conference on High Quality Audio Coding. 1999.

[2] Jürgen Koller. *Improving lossless audio coding*. AES 17th International Conference on High Quality Audio Coding. 1999.

[3] Bernhard Feiten. *Spectral properties of audio signals and masking with aspect to bit data reduction*. AES 86th Convention. 1988.

[4] Detlef Wiese and Gerhard Stoll. *Bitrate reduction of high quality audio signas by modeling the ears masking thresholds.* AES 89th Convention. 1990.

[5] Antonio Pena Giménez. *Técnicas de modelado psicoacústico aplicadas a la codificación de audio de muy alta calidad*. Tesis doctoral. Universidad Politécnica de Madrid. 1994.

[6] Ted Painter and Andreas Spanias. *Perceptual coding of digital audio*. Proceedings of the IEEE, 88(4):449-513. Abril 2000.

[7] G. Stoll et al. *Masking pattern adapted subband coding: use of the dynamic bit-rate margin*. AES 84th Convention. 1988.

[8] Frank Baumgarte et al. *A nonlinear psychoacoustic model applied to the ISO MPEG Layer 3 codec*. AES 99th Convention. 1995.

[9] David J.M. Robinson and Malcom O.J. Hawksford. *Psychoacoustic models and non-linear human hearing*. AES 109th Convention. 2000.

[10] ISO/IEC JTC1/SC29/WG11 MPEG. *International Standard IS 11172-3 Coding of moving pictures and associated audio for digital storage media at up to 1-5 Mbit/s, Part 3: Audio*. 1991.

[11] ISO/IEC JTC1/SC29/WG11 MPEG. *International Standard IS 13818-3 Information Technology – Generic coding of moving pictures and associated audio, part 3: Audio*. 1994.

[12] ISO/IEC JTC1/SC29/WG11 MPEG. *International Standard IS 13818-7 Information Technology – Generic coding of moving pictures and associated audio, Part 7: MPEG-2 AAC*. 1997.

[13] ISO/IEC JTC1/SC29/WG11 14496-3. *Information Technology – Very low bitrate audio-visual coding, Part 3: Audio*. 1998.

[14] ISO/IEC JTC1/SC29/WG11 14496-3 Amd 1/FPDAM. *Information Technology – Coding of audio-visual objects – Part 3: Audio*. 1999.

[15] Bernhard Grill. *The MPEG-4 General Audio Coder.* AES 17th International Conference on High Quality Audio Coding. 1999.

[16] Karlheinz Brandenburg. *MP3 and AAC Explained.* AES 17th International Conference on High Quality Audio Coding. 1999.

[17] Heiko Purnhagen. *An overview of MPEG-4 Audio Version 2*. AES 17th International Conference on High Quality Audio Coding. 1999.

[18] ISO/IEC JTC1/SC29/WG11 N3536. *MPEG-4 Overview – (v.15 – Beijing Version)*. 2000.