UPC

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Contribución a la Regulación del Tráfico en Redes ATM. Aplicación al Tráfico de Vídeo.

TESIS DOCTORAL

Luis J. de la Cruz Llopis
Director: Dr. Jorge Mata Díaz

1999

Tribunal designado por la Comisión de Doctorado de la Universidad Politécnica de Cataluña el día 12 de enero de 1999.

- Presidente Dr. Emilio Sanvicente Gargallo
- Vocal Dr. Alberto González Salvador
- Vocal Dr. Juan Serrat Fernández
- Vocal Dr. Antonio Díaz Estrella
- Secretario Dr. Juan García Haro

Realizado el acto de defensa y lectura de la Tesis Doctoral el día 10 de febrero de 1999.

Calificación:

EL PRESIDENTE

LOS VOCALES

EL SECRETARIO

Dedicado a mi familia
Siempre pienso en vosotros

Agradecimientos

Son numerosas las personas que a lo largo de tanto tiempo de trabajo se han hecho merecedoras de mi más sincero agradecimiento. De entre todas ellas, deseo destacar muy especialmente al director de este trabajo, Jorge Mata, quien ha sobrepasado enormemente las funciones de un director de tesis, trabajando codo con codo desde el inicio hasta el final. Gracias, Jorge, por tu confianza al aceptar dirigir mi trabajo, por tus constantes consejos y soluciones, por tus animadas conversaciones, por tu infinita paciencia y por tu apoyo en los momentos difíciles.

Por otra parte, quiero expresar mi agradecimiento a mi tutor durante el inicio del programa de Doctorado, Emilio Sanvicente, por su apoyo, amabilidad y confianza en mis inicios como profesor asociado.

Este trabajo se ha desarrollado dentro del programa SIGLA, soportado por la Comisión Interministerial de Ciencia Y Tecnología (CICYT) como proyecto TEL 96-1452. Desde aquí quiero expresar también mi agradecimiento al responsable del proyecto, José Luis Melús, por todos los medios que han sido puestos a mi disposición para la elaboración de la Tesis.

También deseo agradecer a mis compañeros del Departamento de Matemática Aplicada y Telemática tanto su ayuda como su colaboración al agradable entorno de trabajo en que ha sido posible desarrollar este trabajo. Entre ellos, Francisco Rico, Joan García, Esteve Pallarès, Mónica Aguilar, Juanjo Alins, Jordi Forga, Francisco Monzó, Marcos Fernández, Jordi Forné y Miguel Soriano destacan por hacer de nuestro encuentro diario algo más que una simple cuestión laboral.

Además, fue de gran ayuda el intercambio de ideas y software de evaluación sobre tráfico autosemejante con Eduardo Casilari y Arcadio Reyes, ambos del Departamento de Tecnología Electrónica de la Universidad de Málaga.

Finalmente, deseo expresar mi agradecimiento a los miembros del tribunal por el desinteresado trabajo realizado en la lectura y evaluación de este trabajo.

Resumen

De entre los nuevos servicios ofrecidos por las actuales redes de banda ancha, los que incluyen la transmisión de secuencias de vídeo se convertirán en un futuro próximo en algunos de los más solicitados por los usuarios. Actualmente, estos servicios son principalmente transportados por canales de tasa binaria constante. Sin embargo, el tráfico de vídeo es por naturaleza variable, con lo que la utilización de recursos de red se vería sustancialmente mejorada utilizando canales de tasa variable.

En este trabajo se plantea el problema de la transmisión eficiente de tráfico de vídeo comprimido MPEG de tasa variable sobre redes ATM de banda ancha. Para ello, en primer lugar se realiza un estudio de los algoritmos de compresión y de los mecanismos de transmisión más comúnmente utilizados en la actualidad. Además, se introducen algunas de las especificaciones que se encargan de estandarizar sistemas y servicios relacionados con la transmisión de vídeo, como el vídeo bajo demanda del ATM Forum o las especificaciones DAVIC.

Posteriormente se lleva a cabo una caracterización del tráfico de vídeo, con objeto de entender la naturaleza del tráfico que se va a entregar a la red para su transmisión. Una de las principales características de este tipo de tráfico es la presencia de dependencias a corto y largo plazo. Estas dependencias son modeladas en este trabajo mediante un filtro ARIMA fraccional, que finalmente se propondrá para la generación artificial de tráfico de vídeo.

Uno de los principales problemas inherentes al algoritmo de codificación MPEG es que la tasa binaria resultante presenta un comportamiento periódico que se repite cada cierto número de imágenes. Este comportamiento es perjudicial tanto cuando se trata de asignar eficientemente recursos de red como cuando se trata de multiplexar distintas secuencias de vídeo. Con objeto de eliminar este comportamiento sin introducir retardos adicionales, se propone un conformador predictivo de tráfico que es comparado con otros esquemas propuestos. La comparación se lleva a cabo por un lado mediante el estudio de los parámetros y funciones que definen la variabilidad de las secuencias estudiadas. Por otra parte, se analizan los recursos necesarios para la transmisión y multiplexación de fuentes de vídeo conformadas y sin conformar.

Finalmente, con objeto de lograr una utilización lo más eficiente posible de los recursos de red, manteniendo una alta calidad de servicio, se propone un esquema de renegociación del ancho de banda asignado. Con este sistema, en momentos de menor actividad se liberarán recursos que podrán ser utilizados por otras conexiones, tanto de servicios de vídeo a tasa variable, como de servicios de tasa constante o de tasa disponible. Por el contrario, en momentos de gran actividad, se podrán solicitar más recursos a la red de forma que el usuario final no perciba ningún cambio en la calidad del servicio que se le está suministrando. El método presentado se basa en la caracterización previa del tráfico de vídeo mediante un modelo de fluidos bidimensional que captura las dependencias a corto y largo plazo del tráfico. Al ser utilizado para la transmisión de un tráfico que previamente ha sido conformado, el modelo debe capturar el comportamiento del tráfico promediado sobre un determinado número de imágenes. Además, permite la detección de los intervalos de distinta actividad, que se han denominado de actividad regular, de gran actividad y de actividad elevada.

Índice

CAPÍTULO 1

INTRODUCCIÓN

Tradicionalmente, las redes de comunicaciones de datos han sido diseñadas con el objetivo de dar servicio a un tipo de tráfico en concreto. Como ejemplos de estas redes se pueden citar las redes telefónicas basadas en conmutación de circuitos para el servicio de telefonía o las redes de conmutación de paquetes para los servicios de transporte de datos. Sin embargo, dada la aparición de nuevos tipos de servicios, y en previsión de los que puedan aparecer, desde hace algunos años se viene trabajando con redes que son capaces de dar servicio a diferentes tipos de tráfico.

Un primer intento de integración se realizó con la Red Digital de Servicios Integrados de Banda Estrecha *(Narrowband Integrated Services Digital Network,* N-ISDN*)*. En el momento de su desarrollo, los servicios más extendidos eran precisamente los de transmisión de datos y de telefonía. Así, esta red se diseñó de forma que permitiese el acceso, a través de un mismo interfaz, a servicios de conmutación de circuitos y a servicios de conmutación de paquetes.

En la actualidad, la aparición de nuevos servicios, como la videotelefonía, la videoconferencia o el vídeo bajo demanda, han hecho insuficientes las velocidades de transmisión que ofrecía la N-ISDN, así como inadecuado su modo de multiplexación. Con el fin de dar un servicio eficiente a estos nuevos tipos de tráfico, y gracias a los avances tanto en la explotación de las líneas de transmisión como en la conmutación fotónica, a finales de los ochenta se comenzó a trabajar en la Red Digital de Servicios Integrados de Banda Ancha (*Broadband ISDN*, B-ISDN). El modo utilizado por esta red para la multiplexación, transmisión y conmutación de la información es el modo de transferencia asíncrono (*Asynchronous Transfer Mode*, ATM). Los puntos claves de la concepción de esta red han sido la integración de servicios y la optimización de la utilización de los recursos disponibles.

1 Motivación

Dentro de las nuevas posibilidades ofrecidas por las redes digitales de servicios integrados de banda ancha destaca la posibilidad del transporte de información con tasa binaria variable. Este transporte se debe llevar a cabo manteniendo una calidad de servicio predeterminada y con una eficiente utilización de los recursos de red disponibles. Entre los servicios de tasa variable que pueden ser soportados por dichas redes, se encuentran los que incluyen información de vídeo comprimido. Estos servicios necesitan un ancho de banda elevado en la red, y se encontrarán en un futuro próximo entre los servicios más solicitados por los usuarios.

El tráfico de vídeo presenta unas especiales características, tanto por su naturaleza como por la calidad de servicio que en ocasiones va a solicitar a la red. Por una parte, se trata de un tráfico muy variable, debido tanto a los algoritmos de codificación utilizados como a la variabilidad de las secuencias a codificar. Además, los servicios que incluyen este tipo de tráfico son sensibles a las pérdidas y al retardo introducido por la red. Así, en general, la calidad de servicio solicitada va a ser alta.

Por parte de la red, la transmisión a tasa constante sería más simple en términos de asignación de recursos, control de admisión de llamadas y control de parámetros de usuario. Esta técnica es la más comúnmente utilizada en la actualidad. Sin embargo, un gran número de trabajos de investigación pone de manifiesto que la utilización de los recursos de red se ve sustancialmente mejorada mediante una transmisión a tasa variable.

Así, la motivación principal de este trabajo es la explotación eficiente de las posibilidades ofrecidas por la B-ISDN para la transmisión de vídeo comprimido MPEG de tasa variable. Para ello, se debe trabajar por una parte sobre el tráfico generado por los codificadores de vídeo, de forma que la entrega de información se haga de la forma más adecuada posible. Por otro lado, será necesario estudiar la estrategia de asignación de recursos por parte de la red con objeto de poder dar un grado de servicio al mayor número de usuarios posible.

2 Objetivos

Este trabajo está centrado en la transmisión de información de vídeo en redes de comunicaciones. Así, como primer objetivo a cumplir se encuentra la revisión de los sistemas y tecnologías que la llevan a cabo. Para ello, se deberán estudiar los sistemas actuales en uso y los que se están desarrollando para un futuro próximo. De esta forma, se podrá tener una amplia visión del problema de la transmisión asíncrona de vídeo y de los estándares sobre los que se trabaja con objeto de garantizar la interoperabilidad entre los distintos fabricantes y operadores.

A continuación, y siguiendo un orden lógico, será necesaria una caracterización lo más completa posible del tráfico con el que se va a trabajar. Este tráfico será la salida a tasa variable de un codificador de vídeo. En la actualidad, son ya numerosos los trabajos de investigación dedicados a este tema, dando como resultado una amplia gama de modelos matemáticos para su caracterización. Estos modelos difieren según las propiedades del tráfico que se tengan en cuenta y de la utilización que se les vaya a dar. Así, el tráfico de vídeo ha sido modelado en diversas escalas de tiempo y teniendo en cuenta varias de sus características más particulares. Entre estas características, la más estudiada en los dos últimos años, así como la más compleja, es su naturaleza fractal. Actualmente, hay abierto un debate sobre la importancia de esta característica a la hora de dimensionar los elementos de red que va a atravesar este tipo de tráfico. En este sentido, este trabajo se centrará en primer lugar en la modelización del tráfico de vídeo a nivel de cuadro teniendo en cuenta precisamente esta última característica. Además, se propondrá un nuevo método de generación sintética de tráfico que incorpore la naturaleza fractal. Como consecuencia, se dispondrá de un conocimiento suficientemente profundo del tráfico sobre el que se va a trabajar que permitirá abordar el tercer objetivo marcado.

A continuación, el trabajo se centrará en el punto de entrega del tráfico de vídeo a la red de comunicación. Para ello será ya necesario conocer los mecanismos de multiplexación, transmisión y conmutación empleados por la red. En este punto, se buscará la forma óptima de entregar el tráfico, de manera que el grado de servicio ofrecido por la red se maximice a través de una eficiente utilización de los recursos de red disponibles. Con este objetivo se utilizarán las caracterizaciones previas llevadas a cabo sobre el tráfico. En este punto se propondrá un interfaz que realice las acciones descritas y se comparará con los existentes.

Finalmente, se dará un paso más adelante en la optimización de la utilización de los recursos de la red. Para ello se trabajará sobre la variabilidad del tráfico entregado a nivel de escena. Como consecuencia, se podrá añadir un nuevo elemento al interfaz comentado que aumentará nuevamente el número de conexiones que podrán ser soportadas con la misma cantidad de recursos de red.

3 Contenido

Como se ha comentado, en este trabajo se plantea la entrega y transmisión eficiente del tráfico de vídeo comprimido MPEG de tasa variable sobre redes ATM. Como introducción, en el capítulo 2 se lleva a cabo una revisión de un sistema de transmisión asíncrona de vídeo, planteando los problemas más importantes a tener en cuenta en su desarrollo. Se introducen los esquemas de codificación más utilizados, haciendo especial hincapié en el algoritmo MPEG al ser el más ampliamente aceptado para el

almacenamiento y transmisión de secuencias de vídeo. Además, se repasan algunas de las especificaciones que actualmente estandarizan varios aspectos de los sistemas y servicios de transmisión de vídeo, como el vídeo bajo demanda del ATM Forum o las especificaciones DAVIC (*Digital Audio-VIsual Council*). En cuanto a la tecnología de transporte, se detallan los conceptos fundamentales de las redes digitales de servicios integrados de banda ancha y del modo de transferencia asíncrono, y se lleva a cabo una breve introducción a las redes de acceso de banda ancha.

A continuación, en el capítulo 3, se procederá a la caracterización del tráfico de vídeo. Entre las características de este tipo de tráfico se encuentra la presencia de dependencias tanto a corto como a largo plazo. Así, se hará un repaso de los conceptos de tráfico autosemejante, de sus propiedades, y de los métodos más usuales de generación y análisis. Además, se propondrá un nuevo método, rápido y sencillo, para la generación de procesos con dependencia a largo plazo. El método es analizado y comparado con otros previamente presentados. Posteriormente, será utilizado también para el ajuste de un nuevo modelo ARIMA fraccional para el tráfico de vídeo a nivel de cuadro. Este modelo será válido para la síntesis de tráfico artificial, tras proyectar la función de distribución gaussiana de la salida generada sobre alguna función de distribución que se adapte mejor a las características del tráfico de vídeo.

En este momento se estará en disposición de estudiar las posibilidades existentes para la entrega del tráfico de vídeo a la red de comunicaciones. Una necesidad fundamental será la de la conformación del tráfico de forma previa a su transmisión. De este modo, en el capítulo 4 se presentarán las técnicas existentes para llevar a cabo dicha conformación, analizando sus ventajas e inconvenientes. Con objeto de mantener las ventajas de la conformación en servicios con fuertes restricciones temporales, se propondrá un nuevo sistema de conformación basado en técnicas de predicción. El método será asimismo analizado y comprobado mediante la simulación de la conformación sobre varias secuencias de vídeo reales. Los resultados serán comparados con los obtenidos con los método clásicos. Posteriormente, se analizarán las ventajas obtenidas mediante la conformación del tráfico en la transmisión sobre redes ATM. Para ello, se presentarán los resultados de simulaciones de transmisión realizadas con trazas de tráfico real conformadas y sin conformar.

Debido a las diferencias en la complejidad de las escenas a codificar, las fuentes de vídeo MPEG VBR pasarán por diferentes niveles de actividad a lo largo de una conexión. De este modo, alcanzar una calidad de servicio determinada pasará por la asignación de los recursos necesarios en los instantes de mayor actividad. Como consecuencia directa, durante la transmisión de las escenas de menor complejidad se estarán desaprovechando los recursos asignados. Por otra parte, los tiempos de permanencia en cada uno de los niveles de actividad serán bastante grandes. Así, en el

capítulo 5, se propondrá la posibilidad de trabajar con una asignación dinámica de recursos a lo largo de la conexión, de forma que la propia fuente de vídeo solicitará o liberará dichos recursos en función de sus necesidades. En este sentido, se propone un nuevo método de implementación del mecanismo de renegociación, basado en la caracterización previa del tráfico a nivel de GoP mediante un modelo de fluidos bidimensional. Gracias a este modelo, se identificarán tres niveles de actividad, y se propondrá un elemento supervisor y controlador, encargado de la solicitud y liberación de recursos en función del nivel de actividad en que se encuentre la fuente. Este elemento será también el encargado de regular la tasa de salida del codificador cuando la red no pueda asignar los recursos deseados por la fuente. De este modo, llevará también a cabo las funciones locales de policía necesarias para no incumplir el contrato establecido con la red. Las ventajas del sistema de renegociación de recursos serán analizadas tanto sobre servicios a tasa constante como sobre servicios a tasa variable.

El objetivo final del trabajo se consigue con la implementación de un interfaz para fuentes de vídeo MPEG de tasa variable sobre redes ATM. El interfaz está basado en la combinación del supervisor controlador con el conformador de tasa presentado en el capítulo anterior. De este modo, mientras el conformador se encarga de alisar el tráfico eliminando las variaciones producidas por el algoritmo de codificación, el supervisor controlador aumenta la eficiencia en la utilización de los canales al solicitar distintas cantidades de recursos en función de la complejidad de las escenas que se estén codificando y transmitiendo. El correcto comportamiento del conjunto de técnicas presentado será validado mediante simulaciones realizadas con secuencias de prueba y con trazas de tráfico real.

Finalmente, en el capítulo 6 se resumen las principales conclusiones extraídas de este trabajo, y se introducen una serie de posibles líneas futuras de investigación.

CAPÍTULO 2

TRANSMISIÓN DE VÍDEO SOBRE REDES ATM

El constante y rápido desarrollo de las técnicas de codificación de vídeo y de las tecnologías de red de banda ancha está provocando la aparición de nuevos servicios cada vez más atractivos para los usuarios. De esta forma, se abre un amplio abanico de posibilidades de negocio, tanto para los fabricantes como para las empresas portadoras y suministradoras de servicios.

La gran cantidad de información generada por las fuentes de vídeo lleva a la necesidad de la aplicación de técnicas de compresión. Esta compresión es posible dada la redundancia, tanto espacial como temporal, presente en dicha información. De entre los algoritmos de codificación de vídeo, el estándar MPEG está siendo el más utilizado tanto para almacenamiento como para transmisión sobre redes de comunicación.

Dentro de las redes de banda ancha, la tecnología ATM se va perfilando como la más adecuada para proporcionar la transmisión de vídeo de la forma más eficiente posible, a la vez que se atienden otros servicios más clásicos como la telefonía o la transmisión de datos. La transferencia de vídeo digital sobre redes ATM ha sido objeto de multitud de trabajos de investigación durante la última década. Uno de los principales retos dentro de este campo ha sido la optimización en el uso de recursos de red manteniendo un grado de servicio determinado. La especial naturaleza de este tipo de tráfico, así como los requisitos de calidad de los servicios que lo incluyen, dan lugar a una serie de aspectos y problemas a tener en cuenta cuando se trata de especificar los sistemas que les darán soporte.

1 Introducción

Las empresas proveedoras de servicios de telecomunciaciones han comenzado ya a proporcionar servicios de banda ancha a sus usuarios además de los clásicos servicios de banda estrecha como la telefonía o la transmisión de datos a baja velocidad. Dentro de estos servicios se encuentran, entre otros, la difusión de vídeo o los accesos a alta velocidad a Internet. Por otra parte, las tecnologías en modems para las líneas de abonado digitales proporcionan la posibilidad de llegar con tasas de bit elevadas hasta los usuarios finales. Las arquitecturas de red de banda ancha, junto con las líneas digitales comentadas para la *última milla*, facilitan la migración entre las redes de banda estrecha y las redes de banda ancha.

Los estándares de audio y vídeo MPEG-2 son los más recientemente adoptados y aceptados internacionalmente para la compresión y transmisión de vídeo y audio digital. Actualmente se utiliza en la mayoría de los sistemas de difusión digital por cable y vía satélite. Entre los organismos que lo han adoptado para su estándares se encuentran el ATM Forum y DAVIC. De ambos se hablará con más detalle posteriormente en este capítulo.

A la hora de transmitir vídeo comprimido MPEG-2 por una red de comunicaciones basada en el modo de transferencia asíncrono, aparecen una serie de aspectos y problemas a tener en cuenta. Por una parte está la calidad de servicio que se debe mantener en una transmisión de vídeo, la cual en general será muy sensible a las pérdidas y retardos introducidos por la red. Las nuevas conexiones aceptadas por la red no deben reducir notablemente la calidad de las ya existentes. Además, cuando el vídeo se transmite comprimido, es muy importante garantizar una probabilidad de pérdidas muy baja, ya que los errores que se produzcan en una imagen pueden verse reproducidos en imágenes posteriores, a pesar de que se introduzcan técnicas de recuperación de errores. Por otro lado, uno de los principales objetivos en el diseño actual de redes de comunicación es la utilización lo más eficiente posible de los recursos disponibles. Por lo tanto, se debe maximizar esta eficiencia para la calidad de servicio requerida por los usuarios.

En este capítulo se expondrán y analizarán los aspectos importantes a tener en cuenta en la transmisión de vídeo digital comprimido MPEG sobre redes de banda ancha basadas en el modo de transferencia asíncrono. Para ello, en primer lugar se describirán los aspectos y generalidades más relevantes a tener en cuenta en un sistema de comunicación de vídeo. Posteriormente se realizará un breve repaso sobre las técnicas de codificación de vídeo, y en particular del estándar MPEG, que será el utilizado a lo largo de todo este trabajo. A continuación se introducirán los conceptos básicos de la Red Digital de Servicios Integrados de Banda Ancha. Con objeto de

entender cómo llega la información hasta el usuario final, se presentarán también algunas de las técnicas utilizadas en las redes de acceso. Finalmente, se expondrán algunas de las especificaciones desarrolladas por diversos organismos para la implementación final de los servicios que incluyen transmisión de vídeo.

2 Transmisión asíncrona de vídeo. Generalidades

La transmisión asíncrona de vídeo se puede definir como la transferencia de señales de vídeo sobre redes que operan mediante multiplexación en tiempo asíncrona, como la B-ISDN. Dicha transferencia se puede llevar a cabo con objeto de visualizar la información en tiempo real o para ser almacenada y ser visionada a posteriori. El primero de estos casos presenta requisitos más estrictos en cuanto a retardo introducido por la red.

En general, un sistema de transmisión digital de vídeo sería como el presentado en la figura 2.1. La señal capturada por la cámara es digitalizada. La información resultante pasa a un codificador donde se aplicará algún proceso de compresión. La secuencia de vídeo es por tanto comprimida y enviada a la red de comunicaciones. En general, el codificador llevará asociado un controlador de tasa que adecuará la transmisión sobre la red. Por su parte, el receptor aplicará el proceso inverso, convirtiendo los datos descomprimidos a una señal analógica que será enviada al monitor para ser visualizada.

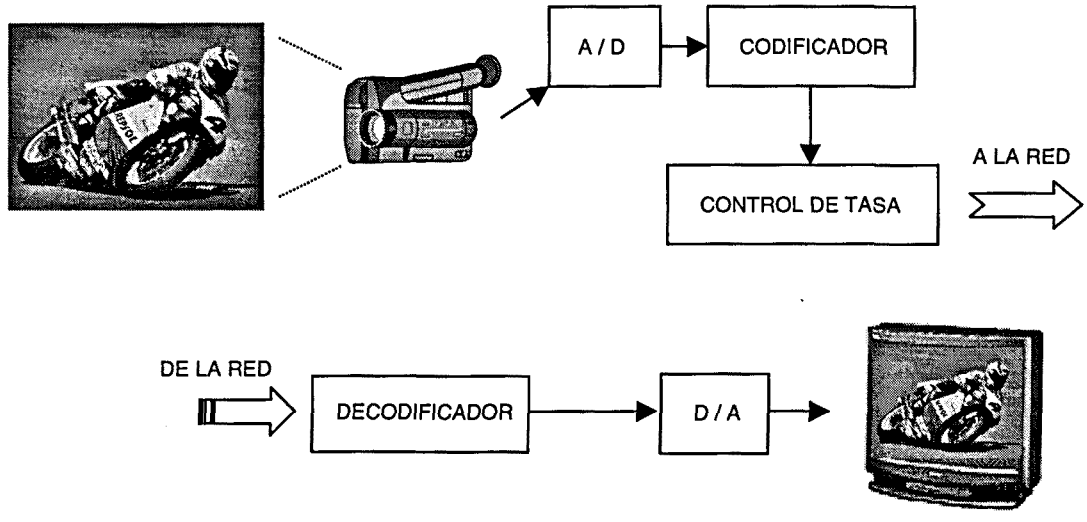

Figura 2.1. Sistema de transmisión de vídeo

El sistema de comunicación debe limitar las pérdidas y el retardo de acuerdo a los requisitos concretos de cada aplicación de vídeo. La mayor parte de la distorsión introducida se produce durante el proceso de codificación. Por otra parte, los errores de transmisión pueden provocar distorsión que se extendería a varias imágenes debido a la dependencia existente entre ellas por el algoritmo de compresión.

Las recomendaciones para el máximo retardo de transferencia siguen las especificadas para las conversaciones de voz. Así, la degradación es mínima por debajo de los 150 ms y muy fuerte por encima de los 400 ms [Kar96]. Por su parte, la información de vídeo puede anteceder a la audio hasta en 100 ms o seguirla hasta en 20 ms. Los valores anteriores son válidos para comunicaciones en un sentido, mientras que para servicios interactivos los requisitos serían aún mayores.

Los usuarios de servicios de comunicaciones incluyendo transmisión de vídeo, esperan la mayor calidad con el coste más reducido posible. Esto lleva a una serie de compromisos a la hora de efectuar la comunicación, esquematizados en la tabla 2.1 [Kar96]. Por ejemplo, la señal de salida de un codificador de vídeo presentará una tasa de bit mayor cuanto más alta deseemos la calidad suministrada. El compromiso existente con el control de tasa será presentado en detalle en el siguiente apartado. Por lo que respecta a la transmisión, al utilizar multiplexaciones estadísticas a lo largo de la red, aparecerán unos retardos de espera en las colas de los distintos nodos de conmutación, que serán tanto mayores cuanto más queramos utilizar y compartir los recursos disponibles. Finalmente, mejorar la fiabilidad del sistema frente a errores de transmisión implica utilizar mecanismos de detección y corrección de errores, que disminuyen el rendimiento del sistema debido a la redundancia (*overhead*) añadida.

	Maximizar	Minimizar
Sesión	Calidad	Coste
Codificación	Calidad	Tasa de bit
Control de tasa	Calidad homogénea	Variabilidad de tasa
Transferencia	Utilización de recursos	Espera en colas
Control de errores	Recuperación de errores	Redundancia añadida

Tabla 2.1. Compromisos en la transmisión asíncrona de vídeo.

2.1 CONTROL DE TASA

Uno de los principales tópicos a la hora de transmitir vídeo comprimido es el control de tasa. Como se verá más adelante, la tasa de bit generada por un codificador de vídeo es variable, debido a dos motivos principales. En primer lugar, la variabilidad de las escenas a codificar, que en unas ocasiones serán más complejas que en otras. Por otro lado, los propios algoritmos de compresión más comúnmente utilizados generan variaciones periódicas de la tasa. Así, para mantener una calidad semiconstante en la transferencia de vídeo, será necesario transmitir una secuencia de tasa variable. Por otra parte, en ocasiones es interesante transmitir a tasa semiconstante, lo cual obligará al codificador a variar la relación de compresión media, provocando asimismo variaciones de calidad. Además, el control de tasa se hará también necesario para evitar que el

codificador genere un número mayor de bits por unidad de tiempo que el que le es permitido transmitir por la red. En la figura 2.2 se esquematizan las dos posibles formas de transmisión, representando con R la tasa de transmisión y con Q la calidad en función del tiempo.

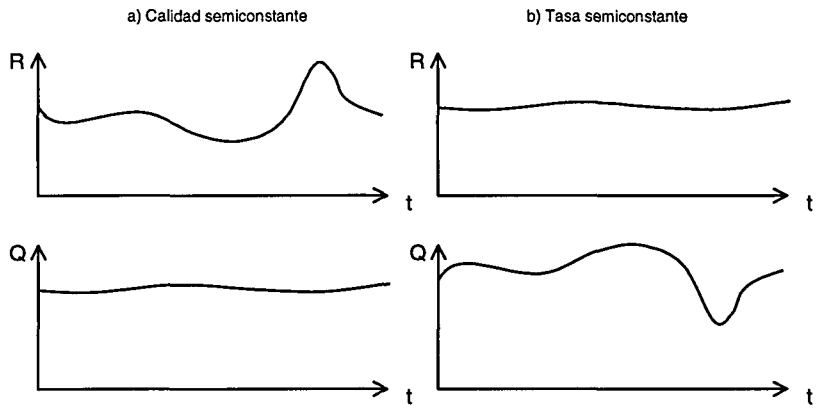

Figura 2.2. Transmisión con tasa variable y con tasa constante.

La codificación a tasa variable (*Variable Bit Rate*, VBR) es común en aplicaciones de almacenamiento, como el disco versátil digital (*Digital Versatile Disc*, DVD). El ahorro de capacidad utilizando esta técnica ha sido analizado en [GriKha98]. Por ejemplo, para almacenar un video-clip de 15 minutos utilizando codificación a tasa constante (*Constant Bit Rate*, CBR) a 6 Mbps se requieren 675 Mbytes, mientras que utilizando VBR con media 3 Mbps y tasa de pico 6 Mbps se necesitan sólo 355 Mbytes. Estos valores se traducen en una ganancia de aproximadamente un 50%, si bien los resultados varían dependiendo de las secuencias a codificar.

En este momento es conveniente distinguir entre codificación VBR y CBR y transmisión VBR y CBR. Por una parte, y como ya se ha comentado, la codificación VBR presenta la gran ventaja de mantener la calidad de la codificación de vídeo constante a lo largo de toda la secuencia. Sin embargo, a la hora de transmitir esta secuencia, puede ser más cómodo para la red que se realice en modo CBR, debido a su mayor simplicidad. El problema de adaptar el flujo VBR proveniente del codificador al canal CBR de la red ha sido propuesto en estándares de transmisión como H.261 y MPEG-1. En varios trabajos de investigación se han propuesto técnicas de conformación tanto a nivel espacial como a nivel temporal. Las primeras tratan de obtener un flujo CBR mediante la multiplexación estadística de varios flujos VBR. Como se comentará posteriormente, en muchos escenarios esta solución no va a ser adecuada. En [LieTse96] y [LieCha97] se propone una técnica denominada *agregación de tráfico*, en la cual se comprimen y multiplexan los distintos flujos de vídeo antes de ser paquetizados y enviados al transporte CBR. Cuando la suma de todos los flujos de vídeo supere la capacidad de transporte CBR disponible se deberá descartar tráfico. La ventaja del sistema reside en que este descarte se lleva a cabo antes de formar los

paquetes, conociendo el tipo de tráfico que se está descartando y consiguiendo que el impacto final en el detrimento de la calidad sea menor. Cuando el descarte se realiza directamente sobre los paquetes, sin importar su contenido, se puede eliminar información muy importante sin la cual enviar otras informaciones no tiene sentido. El principal problema de esta técnica es que sólo es válida en escenarios en los cuales el grupo completo de canales de vídeo deba atravesar conjuntamente toda la red.

Las técnicas de conformación temporal se basan en la inclusión de un buffer entre el codificador y la red, de forma que el tráfico que entra al buffer a tasa variable se va extrayendo de él a tasa constante [LieTse98]. La principal implicación de esta técnica es el aumento del retardo. De este tipo de conformación se tratará con mucho detalle a lo largo de este trabajo.

2.2 TRANSMISIÓN A TASA VARIABLE

Dentro de las transmisiones a tasa variable, es posible hacer una nueva distinción, dependiendo del grado de control que se aplique a dicha tasa. Así, sería posible distinguir entre tasa variable no regulada, suavizada y regulada, como se esquematiza en la figura 2.3. En el primer caso, la información se envía a la red a la misma tasa que es generada por el codificador. En el segundo, un buffer intermedio permite un primer intento de suavizado, pero sin llegar a producir una salida a tasa constante. Finalmente, el tercer caso incluye un control de la tasa binaria dependiendo de la cantidad de información que se vaya acumulando en el buffer, con lo que es posible aumentar o reducir la tasa generada por el codificador. Así, la tasa de salida está más controlada, pero se producirán variaciones de calidad con lo que nos podemos acercar al modelo de tasa constante y calidad variable.

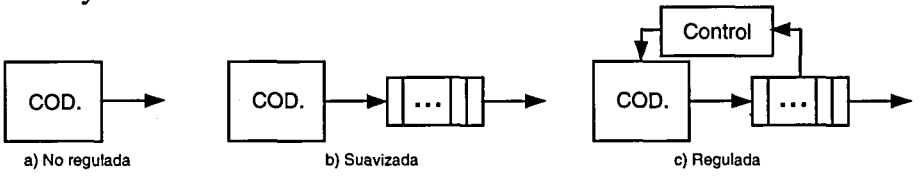

Figura 2.3. Regulación y suavización de tasa variable

La técnica VBR no parece adecuada en primera instancia para transmisión al no limitar la relación de rafagueo de la señal. Posteriormente se hablará con más detalle de este parámetro. Para obtener un ahorro de ancho de banda, como ya se ha comentado, se podrían multiplexar estadísticamente un gran número de flujos de vídeo de forma que el agregado tuviese tasa constante [YegJab93]. De todas formas, esta técnica presenta dos inconvenientes. En primer lugar, el número de fuentes multiplexadas debe ser muy grande, lo cual no va a ser lo más común cuando se habla de fuentes de vídeo. Por otra parte, al transmitir sobre redes conmutadas, los grupos de canales de vídeo no van a ser uniformes a lo largo de ellas. De ambas afirmaciones podemos concluir que, para

aplicaciones de transmisión de vídeo sobre redes conmutadas, la tasa debe ser controlada para cada fuente individual.

Sin embargo, aún controlando de forma individual la tasa generada por las fuentes, los recursos de red serían utilizados de forma más eficiente mediante una transmisión VBR [GriKha98]. En [DalTob97] se lleva a cabo un estudio en base a simulaciones mediante el cual se revela como el número de fuentes VBR multiplexadas puede ser bastante superior al de fuentes CBR, manteniendo los mismos recursos de red y los mismos requisitos de calidad y retardo, especialmente cuando el contenido de la secuencia de vídeo es muy variable. Resultados que corroboran estas afirmaciones se encuentran también en [HsuOrt97][Rei93][PanZar94]. El problema está por una parte en cómo controlar la tasa del codificador, y por otra en cómo asignar los recursos de red de forma más óptima, manteniendo siempre el grado de servicio por encima del nivel esperado en cada aplicación. La finalidad principal de este trabajo de investigación se centra en solucionar, mediante el análisis y la experimentación, las dos cuestiones anteriores.

2.3 RETARDO EXTREMO A EXTREMO

Considerando el sistema global de transmisión de vídeo, la información a transmitir va a sufrir una serie de retardos provocados por cada uno de los procesos a que se ve sometida [Kar98]:

- Adquisición de la imagen
- Codificación y control de tasa
- Segmentación
- Protocolos implicados
- Propagación en los medios de transmisión
- Transmisión en cada uno de los enlaces atravesados
- Espera en colas de multiplexores y conmutadores
- Reensamblado
- Decodificación
- Visualización

Como se ha visto en el apartado anterior, el control de tasa es posible gracias a la inserción de un elemento de almacenamiento entre el codificador y la red. Si el tráfico que se va a enviar a la red está codificado según el algoritmo MPEG, se producen unas fluctuaciones periódicas en la tasa de salida del codificador que también serán suavizadas por este buffer. Sin embargo, el retardo introducido provocará que el sistema no sea válido para servicios con fuertes requisitos temporales. En el capítulo 4 de este trabajo se propone un método de suavizado que reducirá este retardo adicional.

Otro de los factores más importantes de retardo se debe a la segmentación, protocolos de transmisión y posterior reensamblado. Los protocolos incluyen, a su vez, entramado de la información y cálculo de síndromes. Por otro lado, todas estas funciones se deben realizar en nodos que simultáneamente se pueden estar dedicando a otras tareas, lo cual agravaría el problema del retardo. Por ejemplo, el sistema de scheduling de procesos de UNIX puede afectar en gran medida a una transmisión con fuertes requisitos temporales.

Los tiempos de transmisión, por su parte, pueden reducirse aumentando la capacidad de los enlaces y reduciendo el número de nodos intermedios a atravesar. Finalmente, al llevar a cabo un multiplexado asíncrono de la información, aparecerán tiempos de espera en las colas de los elementos de la red.

Por supuesto, es fundamental en servicios en tiempo real interactivos, como por ejemplo el Vídeo Bajo Demanda (*Video on Demand*, VoD), mantener el retardo total por debajo de una cota determinada. Además, las variaciones en el retardo (*jitter*) deben ser ecualizadas en recepción. Esta tarea deberá ser llevada a cabo en el nivel de aplicación, ya que para muchas de ellas no será un problema grave. Además, realizar antes de este nivel la ecualización no garantizaría que no se produzca de nueva una variación en el retardo de los distintos paquetes antes de llegar a la aplicación final.

2.4 ERRORES EN TRANSMISIÓN

Dentro del sistema de comunicación la señal de vídeo está expuesta a diversas distorsiones, como los debidos al ancho de banda limitado de la cámara que captura la imagen o a las cuantificaciones llevadas a cabo por los codificadores.

Dentro del proceso de transmisión existen varias causas de error. Entre ellas:

- Ráfagas de error debidas a ruido en los canales de transmisión
- Pérdidas de paquetes debidas a sobrecarga en las colas de multiplexores y conmutadores
- Errores de encaminamiento
- Retardos por encima de los umbrales permitidos por el servicio

Una de las posibilidades de mejorar el sistema frente a estos errores de transmisión consiste en la utilización de algún código para la detección y corrección de errores. Como se verá más adelante, dentro de este mismo capítulo, un ejemplo puede ser el código de Reed-Solomon (124,128) utilizado por el estándar J.82 de la ITU. Por otra parte, dependiendo del tipo de error que se produzca, en determinadas ocasiones puede ser beneficioso entregar el paquete erróneo al decodificador en lugar de descartarlo por completo. Realmente, la entrega absoluta de información no es necesaria, siendo preferible en determinadas situaciones un pequeño error en un grupo de pixels en la

pantalla que el retardo que supondría la solicitud de una retransmisión del paquete a la fuente.

Por otra parte, en recepción se puede disponer de técnicas de encubrimiento de errores. Una posibilidad es la de rellenar el área para la cual se ha perdido la información con pixels adecuados al entorno, tanto en espacio como en tiempo.

Finalmente, y como se verá con más detalle en el apartado siguiente, hay que tener en cuenta que los errores en un cuadro se pueden propagar a los siguientes.

3 Codificación de vídeo digital

Las recientes aplicaciones y servicios ofrecidos de vídeo han promovido el desarrollo de nuevos algoritmos de comprensión de vídeo digital que reducen sustancialmente la capacidad de almacenamiento y la tasa binaria de transmisión. De entre los posibles servicios ofrecidos cabe destacar los de telefonía [I.F720], videoconferencia [I.F730], distribución de televisión [I.J81], televisión por cable, distribución de televisión de alta resolución [ChiAna94] y vídeo bajo petición [ChaAna94]. El vídeo digital presenta diferentes resoluciones dependientes del servicio o aplicación. Los formatos empleados para los servicios de vídeo parten del formato CCIR-601 [I.R601] especificado para televisión. Así, para videoconferencia y para señal de televisión, con calidad de vídeo doméstico, se emplea el formato CIF (*Common Image Format*) y en servicios de telefonía el QCIF (*Quarter of CIF*). Para compatibilizar la señal de vídeo digital proveniente de vídeo NTSC y PAL también se ha especificado el formato SIF (*Sequence Intermediate Format*) como formato estándar de entrada para los algoritmos de codificación MPEG [MPEG1].

Las técnicas de compresión que emplean los algoritmos para vídeo digital se basan en la explotación de la redundancia espacial y temporal de la señal. El proceso de compresión puede provocar una distorsión o pérdida respecto a la información original, por lo que aparece un compromiso entre el rango de comprensión y la distorsión obtenida. Otras técnicas de compresión no introducen pérdidas pero el rango de compresión resultante suele ser muy inferior.

Las técnicas de compresión se pueden clasificar en función del tipo de explotación de redundancia que realicen. Las técnicas de explotación de la redundancia espacial procesan cada imagen individualmente aprovechando la semejanza entre los pixels de una misma zona, mientras que las técnicas de explotación de la redundancia temporal se basan en el parecido de los pixels situados en una misma posición de un conjunto de campos consecutivos de una secuencia de imágenes.

Las técnicas de compresión basadas en la reducción de la redundancia espacial se pueden clasificar según el tipo de transformación aplicada sobre la imagen en [Ron94]:

- *Codificación predictiva*: Se basa en la codificación del valor diferencial de un pixel respecto al valor estimado a partir de los pixels previamente codificados de su entorno.
- *Codificación transformacional*: Los métodos transformados buscan la extracción de la redundancia de los pixels de una misma zona de la imagen a través de una transformación lineal, de forma que la codificación de los valores obtenidos en el domino transformado sea inferior a la de los pixels de la imagen. Se ha demostrado que la transformación lineal óptima es la denominada transformada Karhunen-Loeve [Kou95]. Esta transformación se basa en que los pixels de una zona próxima están muy correlados y en que la distribución de probabilidad de los pixels de una zona es gaussiana. La transformación óptima se puede aproximar por la transformada discreta coseno (DCT), cuando los coeficientes de correlación están próximos a la unidad. En general, las zonas consideradas de la imagen suelen ser bloques rectangulares de pixels.
- *Codificación en subbandas*: Es una descomposición de la señal original utilizando un banco de filtros de distintas bandas frecuenciales y decimando las señales obtenidas adecuadamente para que no aparezca aliasing. El resultado de este esquema crítico de descomposición en subbandas es un conjunto de señales con un número total de muestras igual a la original. Cada una de las señales se codifica independientemente y se pueden recomponer para obtener diferentes resoluciones de la imagen original.
- *Codificación jerárquica*: La imagen original se descompone en una serie de señales de resolución menor hasta llegar a un nivel básico. A diferencia de la codificación en subbandas, la codificación de cada nivel de resolución necesita de los resultados de la codificación de resolución inferior. En el proceso de decodificación, la imagen original se reconstruye paulatinamente con la agregación de los distintos niveles de resolución.
- *Codificación por segmentación*: Esta codificación se basa en la detección de los contornos de los objetos que componen la imagen y una descripción de estos objetos según su textura, luminosidad, etc. Esta técnica, si bien proporciona elevados niveles de compresión, requiere de un alto coste computacional.
- *Codificación por modelo*: Cuando las imágenes que se pretenden comprimir mantienen invariantes los contornos, como un rostro en videotelefonía, basta con detectar en la imagen aquellos parámetros que describen el objeto invariante y, posteriormente, los correspondientes a su textura. De esta forma, se pueden alcanzar elevados niveles de compresión.

La explotación de la redundancia temporal se realiza a través de las siguientes técnicas:

- *Codificación transformacional*: De la misma forma que se realizaba sobre una zona de una imagen, se puede aplicar la DCT simultáneamente sobre un grupo de pixels situados en diferentes campos consecutivos, pero en la misma zona espacial de cada campo. De esta forma se obtiene la transformación tridimensional denominada 3D DCT.
- *Codificación predictiva*: En este caso, un bloque de pixels se codifica diferencialmente respecto a otro situado en un campo de referencia temporalmente próximo. En general, esta técnica se aplica buscando el bloque de pixels más similar al que se debe codificar, sobre el campo de referencia. Este mecanismo recibe el nombre de compensación de movimiento (CM), de forma que cada bloque codificado predictivamente va unido a un vector de movimiento o desplazamiento relativo del bloque empleado en el cuadro de referencia.
- *Codificación por relleno condicional*: En este caso, en un campo sólo se codifican aquellos pixels cuyo valor es significativamente diferente de los codificados en el campo previo en la misma localización.

Junto con las técnicas de compresión presentadas también se suelen emplear mecanismos de cuantificación. La cuantificación se puede aplicar a cada muestra del dominio de partida o del domino transformado (cuantificación escalar) o sobre un grupo de muestras (cuantificación vectorial) a fin de aprovechar la similitud de muestras próximas.

Los algoritmos de codificación suelen conjugar diversas técnicas de las expuestas anteriormente para maximizar el rango de la compresión para un nivel de distorsión dado o para una tasa binaria constante. En la tabla 2.2 aparecen varios de estos algoritmos junto con los mecanismos de compresión utilizados.

ALGORITMO	CODIFICACIÓN
J.80	Diferencial
H.120	Diferencial y relleno condicional
MJPEG	DCT
H.261	DCT y CM
J.81	DCT y CM
MPEG-1	DCT y CM
MPEG-2	DCT y CM
MPEG-4	Modelo

Tabla 2.2. Algoritmos de compresión de vídeo

4 Codificación de vídeo MPEG-2

Formalmente, el Grupo de Expertos en Imágenes en Movimiento (*Motion Pictures Expert Group*, MPEG) se constituyó en 1988 como *Joint ISO/IEC Technical Committee on Information Technology, Subcommittee 29, Working Group 11 (ISO/IEC JTC1 SC29 WG11)*. Se encargó del desarrollo de estándares para la representación codificada de imágenes en movimiento, la información de audio asociada y su combinación para el almacenamiento en medios digitales [Swe97]. Esta primera fase fue completada en 1991, y de este trabajo surgió la especificación MPEG-1 [MPEG1][Gal91]. Básicamente, el objetivo estaba en conseguir una calidad similar a la de un vídeocasette VHS con una tasa de bit alrededor de los 1.2 Mbps.

La aparición de nuevos servicios en los cuales el vídeo jugaba un papel importante llevó al desarrollo de un nuevo esquema de codificación, basado en el anterior, que abarcase un abanico más amplio de aplicaciones. De este modo se desarrolló un nuevo estándar, MPEG-2 (ISO/IEC 13818) [MPEG2], como un superconjunto de MPEG-1, que adaptaba el anterior a los nuevos servicios emergentes. Originalmente existió también un proyecto MPEG-3 para aplicaciones de televisión de alta definición, pero fue cancelado cuando estas aplicaciones se incluyeron dentro de MPEG-2.

Actualmente, MPEG trabaja en nuevos estándares de codificación para tasa de bit muy bajas. El proyecto es conocido como MPEG-4 [Chi98] y está enfocado en aplicaciones de vídeoconferencia con bajos retardos y requisitos estrictos de ancho de banda. Otro proyecto en desarrollo es MPEG-7, actualmente en la fase conceptual, como intento de estandarizar un conjunto de características que definan los sistemas multimedia [Swe97].

Si bien MPEG-2 se asocia generalmente sólo a la compresión de vídeo, en realidad es una familia de estándares que incluye varios aspectos. En total, son ocho las diferentes partes en que se divide MPEG-2, las cuales se muestran en la tabla 2.3, en la cual se ha respetado la nomenclatura inglesa [OrzSom98].

MPEG-2	DESCRIPCIÓN
ISO/IEC 13818-1	Systems
ISO/IEC 13818-2	Video
ISO/IEC 13818-3	Audio
ISO/IEC 13818-4	Compliance
ISO/IEC 13818-5	Software Simulation
ISO/IEC 13818-6	Digital Storage Media-Command and Control (DSM-CC)
ISO/IEC 13818-9	Real-time Interface for Systems Decoders
ISO/IEC 13818-10	DSM Reference Script Format

Tabla 2.3. Partes del estándar MPEG-2

4.1 CONCEPTOS BÁSICOS

A la hora de definir MPEG-2, uno de los aspectos más importantes era conseguir un alto grado de flexibilidad. Como resultado de ello se permitiría trabajar con distintas resoluciones de vídeo, prestaciones de equipos, requisitos de ancho de banda y calidades de imagen. Así, entre otras características, MPEG-2 no define el método de compresión, sino sólo la cadena de bits resultante. Además, define como se debe decodificar dicha cadena de bits.

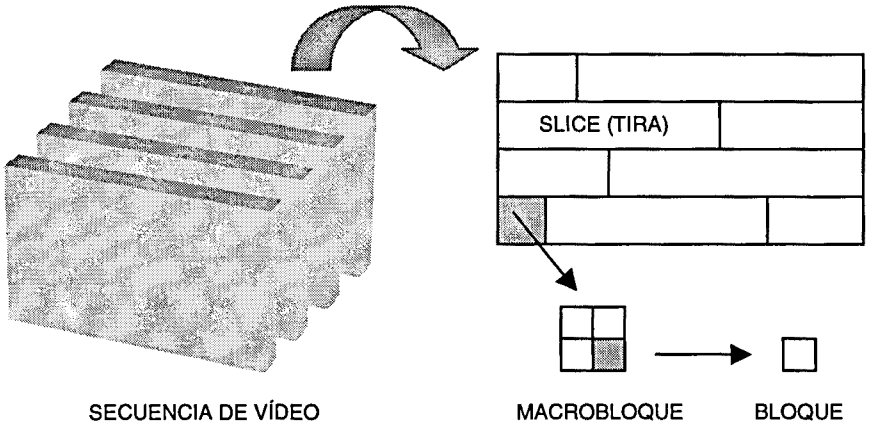

Figura 2.4. Objetos básicos en MPEG-2

En este apartado se introducen los conceptos principales de MPEG, comenzando con los objetos básicos definidos. Dichos objetos se muestran en la figura 2.4. La secuencia SIF se estructura en cuatro niveles de codificación: cuadro, tira o *slice*, macrobloque y bloque. El cuadro es la unidad básica de presentación cuyo número de *pels* (pixels de 8 bits) depende de la resolución. La imagen se estructura en zonas o bloques de 8 x 8 pels donde se aplica la DCT. La agrupación de 4 bloques de luminancia y uno por cada componente de croma se denomina macrobloque. El macrobloque es la unidad básica donde se aplica la técnica de compensación de movimiento. Un conjunto de macrobloques consecutivos horizontalmente se denomina tira o *slice*. La tira es el elemento mínimo donde se puede resincronizar la decodificación en el caso de pérdidas de información. El número de macrobloques consecutivos que forman una tira es seleccionable en el proceso de codificación. En este trabajo se ha considerado la tira como el conjunto de macrobloques que contienen los pixels de 16 líneas consecutivas, es decir, los macrobloques con la misma posición vertical en un cuadro. El presente estudio se ha realizado empleando la resolución estandarizada en la recomendación MPEG-1 de 352 por 288 pels, con submuestreo de las componentes de croma tanto vertical como horizontalmente. Las secuencias analizadas se han obtenido a través de la digitalización de la señal de vídeo PAL o NTSC correspondiente a distintos discos láser (25 y 30 imágenes por segundo respectivamente).

Los cuadros de la secuencia de vídeo pueden codificarse en tres modos diferentes:

- *Intra* (I): son los cuadros codificados empleando únicamente predicción espacial.
- *Predictivo* (P): son los cuadros codificados con predicción temporal hacia atrás, usando como referencia el anterior cuadro I o P, y con predicción espacial
- *Predictivo bidireccional* (B): son los cuadros codificados con compensación de movimiento, empleando como referencias la pasada o futura I o P. La compensación de movimiento se puede realizar sobre los macrobloques de una de las referencias o sobre una semisuma de un macrobloque de cada una ellas.

El almacenamiento o transmisión de las imágenes de una secuencia se hace de forma que el decodificador pueda procesar la información lo antes posible. Para ello, en el almacenamiento o transmisión, las imágenes de referencia preceden a aquellas que las necesitan para ser decodificadas. Este efecto produce en aplicaciones en tiempo real un retardo de reordenación, dado que el orden de decodificación de los cuadros es distinto al de su presentación. A su vez, el codificador también introduce un retardo de proceso dado que necesita imágenes que temporalmente son posteriores para codificar otras que las preceden. Por ello, no es aconsejable en este tipo de aplicaciones que el número de imágenes B consecutivas sea superior a 3 [KawChe93].

La secuencia de imágenes transmitida se estructura en los dos niveles siguientes:

- *Grupo de imágenes* (*Group of Pictures*, GoP, tamaño *N*), compuesto por una imagen I y las imágenes B y P que la han utilizado como referencia.
- *Subgrupo de imágenes* (*Subgroup of Pictures*, SGoP, tamaño *M*) compuesto por una imagen de referencia I o P y las imágenes B que emplearon la imagen I o P como segunda referencia en su proceso de codificación.

En [Mat96] se llevó a cabo un estudio para la elección de los parámetros óptimos N y M. El estudio se basó tanto en la calidad subjetiva obtenida como en un análisis cuantitativo de la misma. La calidad subjetiva se entiende como un nivel de percepción humano en la calidad de la imagen, mientras que la calidad objetiva es una cuantificación que intenta ponderar el error, o distorsión, de la imagen decodificada respecto a la original. En general, la medida empleada en este caso es el PSNR (*Power Signal to Noise Ratio*) [Wan94], definida como:

$$PSNR = 10\log\left(\frac{255^2 R}{\sum_{i=1}^{R}(p'(i)-p(i))^2}\right) \quad [dB] \tag{2.1}$$

donde R es el número de pels en la porción de imagen a analizar, $p(i)$ es el valor del pel original y $p'(i)$ es el valor del pel decodificado. Ambos enfoques llegaron a conclusiones similares, proponiendo los valores N=4 y M=2 o N=6 y M=2. Estos valores han sido

también los adoptados en la mayoría de secuencias de prueba utilizadas en este estudio. En la figura 2.5 se ilustra la secuencia de imágenes para los valores N=6 y M=2, indicando su orden de transmisión y de visualización.

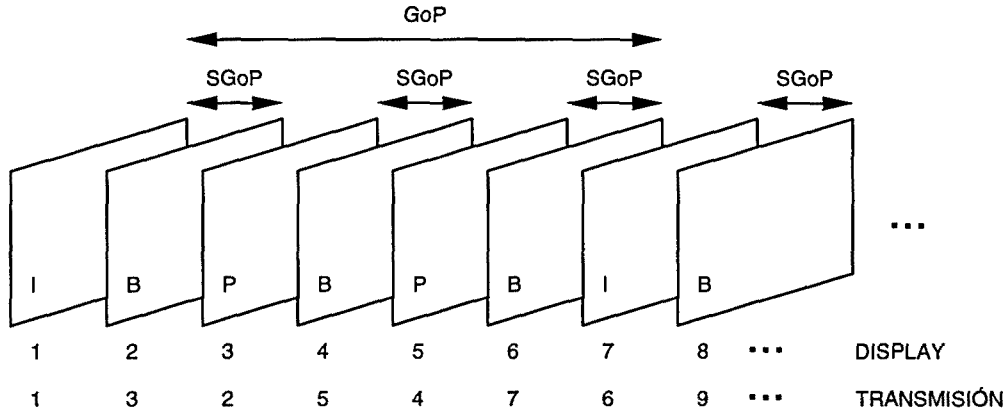

Figura 2.5. Estructura de la secuencia de imágenes (N=6, M=2)

El algoritmo de codificación MPEG tiene dos modos de operación, configurables según el tipo de aplicación para la cual se emplea la compresión. En transmisiones sobre circuitos de capacidad fija, el algoritmo se configura para generar una tasa binaria constante, modo CBR. En el caso de que el sistema de comunicaciones soporte servicios de tasa variable, el algoritmo se puede configurar en modo VBR. El modo de funcionamiento VBR presenta la ventaja, respecto al CBR, de poder mantener una calidad, subjetiva u objetiva, constante en toda la secuencia codificada de imágenes [SimRos93]. Además, la utilización de recursos de red es más eficiente [GriKha98].

Las variaciones de la tasa binaria generada en la codificación se deben a razones tanto intrínsecas, debidas al algoritmo de codificación, como extrínsecas debidas a la complejidad y actividad de la secuencia a codificar. Las razones intrínsecas están relacionadas, fundamentalmente, con los modos de codificación aplicados sobre las imágenes. Así, las imágenes I necesitan un número superior de bits a las imágenes P o B al emplear únicamente la técnica transformada DCT. Asimismo, las imágenes P suelen generar mayor número de bits que las B, dado que sólo emplean compensación de movimiento respecto a las imágenes de referencia anteriores. Dentro de la codificación de las imágenes, otro factor que provoca variaciones de la tasa binaria generada es la explotación de la entropía a través de tablas de códigos de longitud variable, según el tamaño y posición de las ráfagas (run-length) de los coeficientes de la DCT.

Las razones extrínsecas que provocan variaciones en la tasa binaria dependen del contenido de las imágenes a codificar. Las imágenes con mayor grado de detalle o mayor relieve tienen un nivel de complejidad superior y reducen la efectividad de la explotación de la redundancia espacial. Las secuencias de gran actividad, con movimientos rápidos de cámara, zooms y cambios de plano, impiden el empleo de la

técnica de compresión predictiva, por lo que, también provocan aumentos en la tasa binaria respecto a secuencias de menor actividad.

Para conseguir una tasa binaria constante en el modo de funcionamiento CBR es preciso intercalar entre la salida del codificador y el canal, una memoria tampón, o buffer, que absorba las variaciones de la tasa binaria generada en la codificación. Los bits almacenados en el buffer son extraídos a velocidad constante, mientras que el codificador llena el contenido del buffer de forma irregular. Para controlar el retardo introducido por la inserción del buffer, se dimensiona éste con una capacidad limitada y se regula la tasa de generación en el proceso de codificación dependiendo del nivel de ocupación del buffer.

La tasa de generación y el nivel de distorsión de la imagen se pueden controlar a través de varios parámetros que intervienen en el proceso de codificación. Los parámetros controlables que afectan a la generación de la tasa binaria son la resolución espacial y temporal de la secuencia, el número N o M de imágenes que componen un GoP o un SGoP, la cuantificación de las tiras de imagen o la cuantificación de los macrobloques individualmente. Los parámetros no controlables son los extrínsecos como el contenido estadístico de la secuencia y la actividad de la escena. En general, los parámetros de resolución y frecuencia de las imágenes se acuerdan al principio de la sesión y no se modifican en su transcurso.

El parámetro más adecuado para controlar la tasa de generación o el nivel de calidad de la imagen es el paso de cuantificación para un macrobloque o para el conjunto de macrobloques de una tira de imagen. Es el más apropiado dado que no introduce una sustancial sobrecarga de señalización y tiene una rápida respuesta temporal sin añadir un elevado coste computacional. También se puede utilizar, como parámetro de control, la variación del número de elementos que componen un GoP o SGoP. Este mecanismo no se puede emplear de forma sostenida cuando aumenta la complejidad de la secuencia, ya que el exceso de imágenes B provoca automáticamente un aumento de la tasa binaria debido al incremento de macrobloques codificados en modo intra.

La compresión en MPEG-2 se consigue en base a tres técnicas. En primer lugar se extrae la redundancia espacial mediante la cuantificación de los coeficientes obtenidos tras aplicar la DCT. Esta cuantificación se lleva a cabo dividiendo por unos factores proporcionados por una matriz de ponderación. Además, está controlada por un factor de escala que permite al usuario ajustar el nivel de compresión. Tras este proceso, se lleva a cabo un barrido en zig-zag del plano de coeficientes DCT, con lo que las altas frecuencias quedan agrupadas al final. Esto provocará un gran número de ceros, lo cual será aprovechado en el siguiente paso.

En segundo lugar, se realiza una codificación de Huffman de los coeficientes. Este tipo de codificación asigna palabras código de mayor número de bits a los coeficientes menos comunes, reservando las de menor número de bits para los coeficientes más usuales (*Variable Length Coding*, VLC).

Finalmente, se utiliza la técnica de compensación de movimiento para los cuadros B y P. Se buscan macrobloques parecidos en el cuadro actual y en el de referencia y se obtiene la diferencia. Dicho valor se transforma mediante la DCT y posteriormente se codifica VLC junto con el vector de movimiento del macrobloque. En el mejor de los casos, un macrobloque se repetirá exactamente de la misma forma en el cuadro actual y en el de referencia, y además permanecerá en la misma posición. De esta forma se tendrá una diferencia y un vector de movimiento nulos.

La estructura sintáctica presentada por MPEG-2 Video es muy variable debido a los diferentes perfiles y aplicaciones. Así, la aparición de algunos campos depende del valor que adopten otros. En otras ocasiones, la aparición o no de algún elemento depende del valor tomado por un bit (*flag*) dentro de otro elemento de control. De esta forma se consigue además reducir la cantidad de bits a transmitir, evitándose la transmisión de ceros o de valores nulos que no aportan información.

4.2 ESCALABILIDAD

El estándar de codificación MPEG-2 amplía las aplicaciones a las que estaba dirigido el MPEG-1. Las principales mejoras introducidas en el MPEG-2 son: la posibilidad de operar con imágenes entrelazadas al emplear compensación de movimiento sobre macrobloques de 16 x 8 pels, aumenta la precisión de los coeficientes de continua de la DCT a 10 bits frente a los 8 de MPEG-1, permite la cuantificación no lineal, mejora el control frente a errores en su sintaxis e introduce el concepto de escalabilidad.

Esta última es una de las características más importantes de MPEG-2 Video, proporcionando soporte para un amplio rango de aplicaciones de vídeo. MPEG-2 se puede utilizar para distribución estándar de TV, para TV de alta definición (*High Definition TV*, HDTV), o para la transmisión de imágenes de vídeo a través de redes de telecomunicación.

Para conseguir la escalabilidad la información de vídeo se separa en diferentes flujos o niveles de información, los cuales son complementarios entre sí. Una aplicación básica sería tener un flujo o nivel base para una transmisión de TV estándar (PAL o NTSC), al que se podría añadir un nivel de mejora conteniendo información adicional para proporcionar una transmisión del mismo programa en HDTV. Dependiendo de las características del receptor, se quedaría sólo con el nivel básico o bien decodificaría los dos niveles.

El estándar MPEG-2 define varios modos de escalabilidad:

- *Escalabilidad espacial*: Capacidad para trabajar con diferentes resoluciones de pantalla. En este caso, los niveles básico y de mejora se combinan tras realizar la DCT inversa.
- *Escalabilidad temporal*: Se define como la posibilidad de manejar diferentes tasas de cuadro en un mismo flujo de vídeo. El nivel base, proporcionando una tasa estándar, se puede combinar con el nivel de mejora para alcanzar mayores tasas de cuadro. Los niveles básico y de mejora se combinan, al igual que en el caso anterior, tras realizar la DCT inversa.
- *Escalabilidad en SNR*: Permite manejar al menos dos calidades de vídeo diferentes. La información proporcionada por el nivel base puede ser realzada por dos o más niveles de mejora. Sin embargo, todos los niveles tienen la misma resolución espacial. La principal aplicación es en el encubrimiento de errores. Así, el nivel base podría transportar la información más crítica utilizando un canal más robusto, mientras el nivel de mejora es transmitido por un canal menos fiable. Los errores en este canal de mejora no se harán tan patentes durante la decodificación, ya que al menos será posible presentar las imágenes proporcionadas por el nivel base.
- *Partición de datos*: Este proceso se utiliza para dividir el flujo de información en partes más y menos importantes. De nuevo, las parte más importante, en este caso de la sintaxis del flujo de vídeo MPEG-2, se envía por un canal más fiable mientras que la menos importante se puede transmitir por un medio menos robusto. Una posibilidad sería enviar los elementos sintácticos de alto nivel (como las cabeceras de la secuencia, de los GoPs o de los cuadros), junto con el primer coeficiente de la DCT por el canal básico, mientras que el resto de coeficientes de la DCT se transmitirían por el canal de mejora. Existe un elemento especial, *priority breakpoint*, que define qué partes del flujo de vídeo se ponen en cada partición.

4.3 PERFILES Y NIVELES

El gran rango de aplicaciones al que va dirigido MPEG-2 Video tiene como consecuencia un estándar complejo. En muchas ocasiones, los servicios ofrecidos a los usuarios no necesitarán hacer uso de gran parte de las posibilidades ofrecidas. Por otro lado, no tiene sentido en estos casos que los decodificadores sean capaces de entender todas las posibilidades de MPEG-2, lo cual los hace más complejos y encarece, si el usuario final no va a poder sacar partido de ellos.

Con objeto de flexibilizar la utilización de MPEG-2 Video se definen por tanto una serie de perfiles (*profiles*) y niveles (*levels*) formados por subconjuntos de las

posibilidades ofrecidas por el estándar completo. Estos perfiles y niveles aparecen en la tabla 2.4.

PROFILES	LEVELS
Simple Profile (SP)	Low Level (LL)
Main Profile (MP)	Main Level (ML)
SNR Scaleable Profile	High 1440 Level (H14)
Spatial Scaleable Profile	High Level (HL)
High Profile	

Tabla 2.4. Perfiles y niveles de MPEG-2 Video

Un perfil se describe como un subconjunto bien definido de la sintaxis de vídeo. Algunos elementos del estándar no serán válidos ni podrán ser decodificados si el decodificador sólo proporciona un perfil bajo. Por ejemplo, el perfil simple (SP) no admite cuadros B. Los perfiles simple y principal (SP y MP) no soportan ningún tipo de escalabilidad. Los perfiles bajos son siempre subconjuntos de los más altos. En la tabla 2.5 se presentan algunas de las características de los perfiles comentados.

Facilidad MPEG-2	Simple Profile	Main Profile	SNR Profile	Spatial Profile	High Profile
Formato Croma	4:2:0	4:2:0	4:2:0	4:2:0	4:2:0 ó 4:2:2
Cuadros B	No	Si	Si	Si	Si
Modo esc.	Ninguno	Ninguno	SNR	SNR o Espacial	SNR o Espacial

Tabla 2.5. Requisitos para perfiles MPEG-2

Por otro lado, los niveles definen valores para ciertos parámetros, como por ejemplo el número de líneas por cuadro o el número de cuadros por segundo. Perfiles y niveles son combinados para definir exactamente qué subconjunto de MPEG-2 Video se está utilizando. Una combinación muy importante es la conocida como *"Main Level at Main Profile"* (ML@MP). Esta combinación es adecuada para la difusión de TV, con calidad PAL o NTSC. Algunos de los parámetros aparecen en la tabla 2.6.

PARÁMETRO	VALOR EN ML@MP
Muestras por línea	720
Líneas por cuadro	576
Cuadros por segundo	30
Muestras de luminancia por segundo	10368000
Tasa máxima de vídeo	15 Mbps
Tamaño máximo de buffer del decodificador	1835008 bits

Tabla 2.6. Valores MPEG-2 ML@MP

4.4 MPEG-2 AUDIO

La especificación de audio de MPEG-2 [MPEG2_3] es una extensión de la que existía en MPEG-1, presentando un alto grado de compatibilidad con esta, hasta el punto de que un decodificador audio MPEG-1 es capaz de decodificar parte de la información codificada en MPEG-2. Ambos estándares describen tres niveles de compresión, aumentando tanto la compresión como la calidad al pasar del nivel 1 al 2 y del 2 al 3. Los tres niveles son compatibles, en el sentido de que un decodificador de nivel *N* es capaz de decodificar la información del nivel *N-1*.

La aparición de tres niveles es en gran parte una consecuencia histórica, ya que la especificación de nivel 3 es posterior a las anteriores, con lo que estas dos tenían ganada una amplia cuota de mercado, que se mantiene en la actualidad. Sin embargo, la excelente capacidad de compresión y la gran calidad suministrada, hacen de MPEG-2 Audio nivel 3 la mejor elección en la mayoría de las aplicaciones actuales. Su extensión actual es muy grande, con infinidad de servidores web dedicados exclusivamente a la comercialización de temas musicales codificados en este formato. La ventaja es clara: en un CD-ROM clásico es posible almacenar del orden de 170 canciones comprimidas, las cuales son decodificadas en tiempo real sin problemas en un PC doméstico, ofreciendo una calidad de sonido similar a la de un disco compacto. En la tabla 2.7 se resumen las principales características de los tres niveles comentados.

Nivel	Compresión aproximada	Margen de tasa de bit	Retardo mínimo teórico
1	1:4	32-448 Kbps	19 mseg.
2	1:6	32-384 Kbps	35 mseg.
3	1:10	32:320 Kbps.	58 mseg.

Tabla 2.7. Niveles de codificación MPEG-2 Audio

Como se ha comentado MPEG-2 Audio toma como base MPEG-1 Audio. Algunas de las diferencias y mejoras se enumeran a continuación:

- *Frecuencia de muestreo reducida*: En MPEG-2 es posible utilizar una tasa de muestreo reducida a la mitad, y continuar obteniendo una buena calidad de sonido.
- *Extensión multicanal*: Con objeto de obtener una representación estereofónica más realista, se habilitan cinco canales de audio que proporcionan una audición estereofónica "envolvente" (*surround*). Los cinco canales se conocen como izquierdo (*Left*, L), derecho (*Right*, R), central (*Center*, C), envolvente trasero izquierdo (*Left rear Surround*, LS) y envolvente trasero derecho (*Right rear Surround*, RS). Además, se puede incluir un canal especial de baja frecuencia (*Low Frequency Enhancement*, LFE), entre 15 y 120 Hz, principalmente dedicado a la reproducción de efectos especiales.

La compatibilidad comentada entre MPEG-1 y MPEG-2 obliga a ciertas acciones que no permiten alcanzar la máxima calidad posible con la misma tasa de compresión. Como consecuencia, se ha formado un grupo que no respeta la compatibilidad con esquemas anteriores (*Non-Backward Compatible*, NBC), con el objetivo de obtener calidades de sonido superiores a tasas de bit equivalentes.

4.5 MPEG-2 SISTEMAS

La misión principal de MPEG-2 Systems [MPEG2_2] es la de proporcionar una especificación genérica para la multiplexación conjunta de la información codificada de audio y vídeo, independiente de la red física por la que se transmita. Así, esta parte de la especificación MPEG puede considerarse como el interfaz entre los codificadores de audio y vídeo por un lado, y la red de comunicaciones por otro. En la figura 2.6 se esquematiza esta relación.

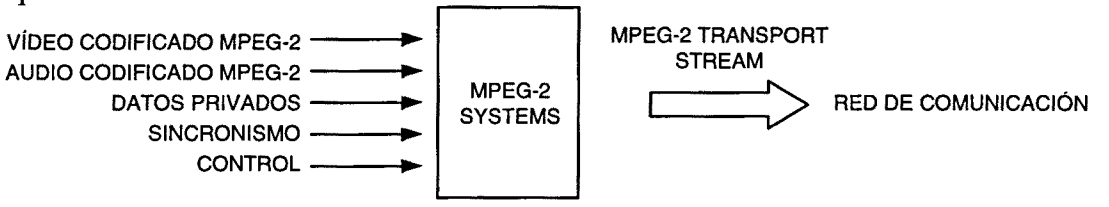

Figura 2.6. MPEG-2 Systems básico

MPEG-2 Systems distingue entre flujo de programa (*Program Stream*, PS) o flujo de transporte (*Transport Stream*, TS), en función de si la información entregada va a ser almacenada para una posterior visualización, o si por el contrario va a ser transmitida por una red de comunicaciones. Por tanto, para el propósito que nos ocupa, nos centraremos en el TS. En este caso, las estructuras de datos utilizadas son cortas y de longitud fija para facilitar su envío por la red.

Una de las misiones fundamentales será la de multiplexar y demultiplexar distintos programas transportando diversos flujos de audio y vídeo. La sincronización necesaria se lleva a cabo añadiendo marcas temporales (*timestamps*). Además, también es posible añadir datos a la transmisión, así como cierta información de control y de gestión. Toda esta información es multiplexada en el TS.

Los diferentes elementos de información manejados se muestran en la figura 2.7 y se explican a continuación. En MPEG-2, el flujo de salida de un codificador de audio o vídeo se conoce como flujo elemental (*Elementary Stream,* ES). Como se ha comentado anteriormente, existe la posibilidad de añadir datos privados a la comunicación, que también formarían un flujo elemental. El flujo elemental se divide en unidades de acceso, que para el caso del vídeo estarían formadas por las distintas imágenes (I, P, B) a transmitir, como se observa en la parte superior de la figura. Este flujo elemental se paquetiza (*Packetized Elementary Stream*, PES), siendo el tamaño del paquete variable

y conteniendo exactamente una unidad de acceso. Posteriormente, los paquetes PES se mapean dentro de los paquetes de flujo de transporte MPEG-2 (*Transport Stream Packets*, TSP). Dichos paquetes, de tamaño fijo, forman el MPEG-2 TS. El motivo de esta doble paquetización es el de crear dos niveles con distintos objetivos. Mientras las cabeceras PES contienen información directamente relacionada con el ES, como por ejemplo si se trata de audio, de vídeo o de datos, las cabeceras TS transportan información útil para la transferencia y entrega del flujo de información.

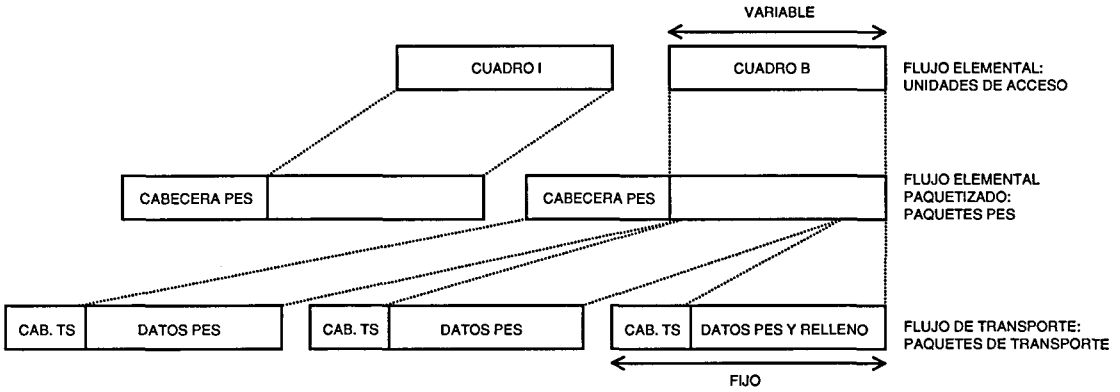

Figura 2.7. Relación entre unidades de acceso, paquetes PES y paquetes de transporte

Por otra parte, se distinguen dos tipos de flujo de transporte: *Single Program Transport Stream* (SPTS) y *Multiple Program Transport Stream* (MPTS). El SPTS contiene diferentes flujos PES, los cuales comparten una base de tiempos común. Por su parte, el MPTS multiplexa varios SPTSs, dando lugar a una jerarquía como la mostrada en la figura 2.8.

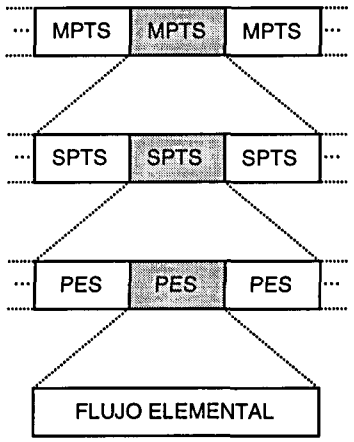

Figura 2.8. Jerarquía en MPEG-2 Systems

5 Red digital de servicios integrados de banda ancha

Con objeto de llevar a cabo la transmisión de señales de vídeo digital comprimido MPEG, se hace necesario un adecuado mecanismo de transmisión. Actualmente, la tecnología que mejor cubre las necesidades en cuanto a ancho de banda, flexibilidad e interactividad de los servicios de vídeo digital es la proporcionada por la Red Digital de

Servicios Integrados de Banda Ancha (*Broadband Integrated Service Digital Network*, B-ISDN) [Pry93][Sch96][Gar95]. La ITU normalizó el modo de transferencia asíncrono (*Asynchronous Transfer Mode*, ATM) como mecanismo de multiplexado, transmisión y conmutación para este tipo de redes [I.113][I.121]. Esta tecnología ha sido adoptada para la transmisión de vídeo junto a otros servicios en redes residenciales de banda ancha por compañías como British Telecom, Deutshe Telekom, Hong Kong Telecom, NTT y US West entre otras [Wri97]. Además, se utiliza con propósitos tan distintos como la educación y los negocios, así como con fines militares. En la actualidad, son ya numerosas las recomendaciones propuestas tanto por los comités de desarrollo de la ITU-T como por el ATM Forum. Ésta última es una organización internacional sin ánimo de lucro formada con el objetivo de acelerar el uso de los productos y de los servicios ATM a través de una rápida convergencia en las especificaciones de interoperabilidad. Fundado en 1991, actualmente cuenta con alrededor de 600 compañías miembro. En este apartado se llevará a cabo un repaso de los conceptos más básicos de la B-ISDN y del ATM.

La B-ISDN se caracteriza básicamente por la habilidad de dar servicio a todos los tipos de información presentes y futuros, a velocidades de transmisión altas y de la forma más eficiente posible [Pry93]. Esto contrasta con la situación más reciente, en la cual diferentes tipos de red se mantienen en funcionamiento para dar servicio a distintas clases de tráfico. La actual estructura, en la cual las redes telefónicas conmutadas proporcionan servicios de telefonía y de datos a baja velocidad, mientras que las conexiones X.25 o Frame Relay están dedicadas a la transmisión de datos a velocidad media, presentan el principal inconveniente de que son redes distintas que deben ser instaladas y gestionadas de forma separada. Además, si una de ellas está muy libre de tráfico en un momento dado, sus recursos no pueden ser dedicados a otros tipos de servicios, con lo que se lleva a cabo una utilización ineficiente de los recursos disponibles.

Un primer intento de unificación de redes se llevó a cabo con la Red Digital de Servicios Integrados de Banda Estrecha (*Narrowband* ISDN, N-ISDN). Esta red permite la transmisión de voz y datos de forma conjunta. Sin embargo, presenta una serie de limitaciones. En primer lugar, el ancho de banda es limitado con lo que no es adecuada para la transmisión de servicios de distribución de vídeo, ni siquiera comprimido. Además, es una tecnología inflexible en varios aspectos, como por ejemplo en la asignación estática de ancho de banda a las conexiones.

De esta forma, la B-ISDN aparece como la nueva red de servicios integrados capaz de proporcionar servicio a los más diversos tipos de información conocidos, así como a los que puedan aparecer en el futuro. Además, lleva a cabo una utilización más eficiente del ancho de banda disponible gracias al empleo del modo de transferencia asíncrono.

Otro de los aspectos innovadores es la posibilidad de la contratación de diversos grados de calidad de servicio (*Quality of Service*, QoS) en el momento del establecimiento de la conexión.

5.1 PROTOCOLOS B-ISDN

Al igual que ocurre con otros sistemas de comunicación, los protocolos utilizados en la B-ISDN siguen un modelo de niveles. La figura 2.9 muestra el modelo de referencia de protocolos, mientras que la figura 2.10 presenta las funciones de cada capa según se describe en [I.321][I.413].

El modelo de referencia de protocolos consta de tres planos: el plano de usuario, el plano de control y el plano de gestión. El plano de usuario hace referencia a la transferencia de información de usuario e incluye mecanismos como control de flujo y recuperación de errores. El plano de control se encarga del control de las llamadas y conexiones. Es decir, comporta todas aquellas funciones de señalización necesarias para establecer, supervisar y liberar las conexiones.

El plano de gestión es el responsable de funciones de operación y mantenimiento (*Organization and Management*, OAM) de la red. La gestión de planos realiza funciones de gestión relativas al sistema global y proporciona la coordinación necesaria entre planos. No posee una estructura de capas. La gestión de capas realiza funciones de gestión que tienen que ver con los recursos y parámetros que residen en sus entidades de protocolo. También maneja flujos de información de operación y mantenimiento, pero específicos a cada capa en cuestión.

La capa física (*Physical Layer*, PL) se divide en dos subcapas: la subcapa del medio físico (*Physical Medium*, PM) y la subcapa de convergencia a la transmisión (*Transmission Convergence*, TC) [I.432]. La subcapa PM incluye todas aquellas funciones dependientes del medio físico en cuestión (p.e., conversión electro-óptica). La subcapa TC efectúa todas las funciones necesarias para transformar un flujo de celdas en un flujo de entidades de datos (por ejemplo bits), compatible con el esquema de multiplexación del sistema de transmisión.

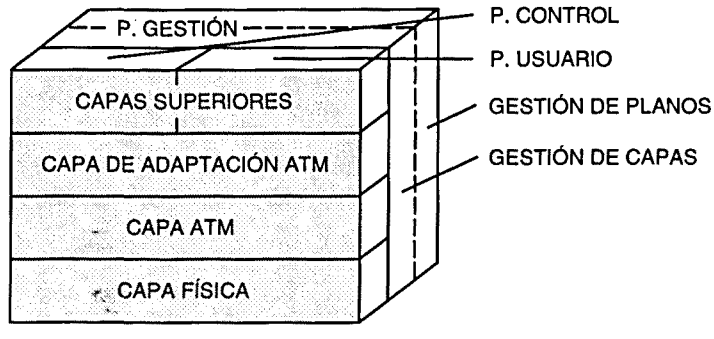

Figura 2.9. Modelo de referencia de protocolos para la RDSI-BA

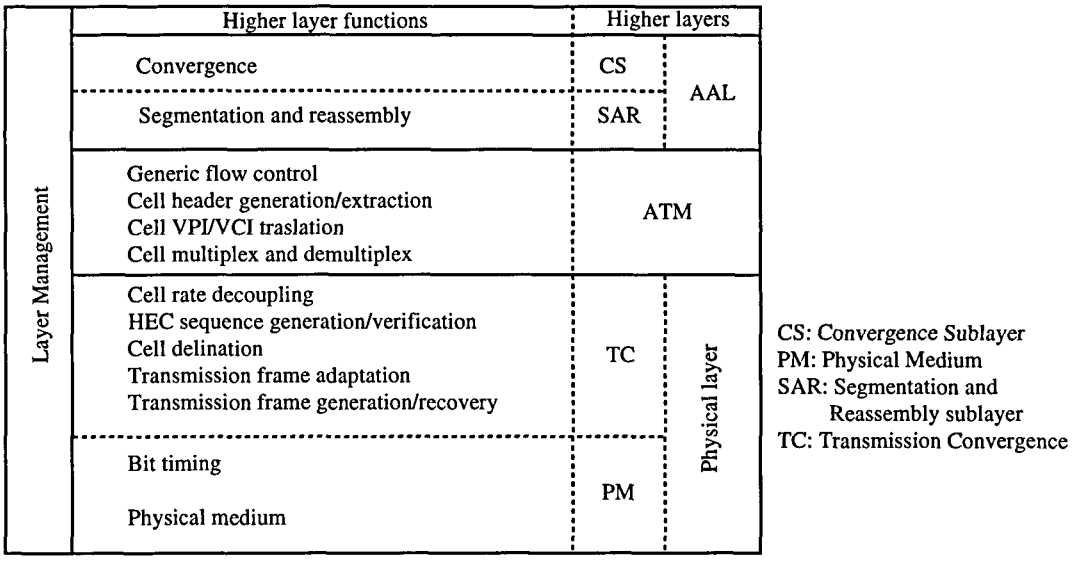

Figura 2.10. Funciones de la RDSI-BA en relación al modelo de referencia

La capa de adaptación ATM (*ATM Adaptation Layer*, AAL) [I.362] [I.363] tiene como función básica el aislar las capas superiores de las características propias de la capa ATM. Se encarga de adaptar los datos procedentes de niveles superiores a un formato que pueda ser manipulado por la capa ATM. La capa AAL se organiza en dos subcapas. La subcapa de convergencia (*Convergence Sublayer*, CS) y la subcapa de segmentación y reensamblado (*Segmentation And Reassembly*, SAR). La subcapa SAR tiene como misión el segmentar los datos de las capas superiores a un formato compatible con el campo de información de usuario de una celda ATM (48 octetos), o recíprocamente, el reensamblar dichos campos de información a unidades de datos de protocolo (*Protocol Data Unit*, PDU) de la capa superior.

La subcapa CS es independiente del servicio y realiza funciones necesarias para soportar aplicaciones específicas (p.e., recuperación de reloj en servicios de vídeo insertando palabras de sincronización). Los posibles mecanismos de adaptación definidos para el transporte de información en redes ATM [I.362] se han planteado según el tipo de servicio que deben soportar. Los parámetros que clasifican un tipo de servicio son [Pry93]:

- Necesidad de relación temporal entre fuente y destino.
- Tasa generada.
- Modo de conexión.

A partir de estos parámetros se definen cuatro tipos de servicios [I.211]:

- Servicios de Clase A. Tienen requerimientos temporales, su tasa binaria es constante y están orientados a conexión. Para estos servicios la red ATM establece una conexión que emula un circuito.

- Servicios de Clase B. Tienen las mismas características que los de clase A salvo que la tasa generada es variable. Estos son los servicios clásicos de vídeo y audio comprimidos con calidad semi-constante.
- Servicios de Clase C. No tienen requerimientos temporales, su tasa es variable y están orientados a conexión. Estos servicios son clásicos en la conexión entre redes locales.
- Servicios de Clase D. Son servicios de tasa variable, sin requerimientos temporales y no orientados a conexión. Son servicios de tráfico de datos no orientados a la conexión.

Las distintas capas AAL definidas se corresponden con los diferentes servicios de la siguiente forma:

- AAL-1: definida para soportar aplicaciones de clase A.
- AAL-2: definida para soportar aplicaciones de clase B.
- AAL-3/4: cuando se aceptó que un sólo protocolo AAL podía usarse para soportar servicios de datos orientados a conexión y en modo datagrama (no orientados a conexión) se especificó la capa AAL-3/4 para tratar ambos servicios, o sea, servicios clase C y D.
- AAL-5: como resultado de la complejidad asociada con la capa AAL-3/4, se propuso el nivel AAL-5, también conocido como capa de adaptación simple y eficiente (*Simple and Efficient Adaptation Layer*, SEAL). Proporciona funciones más limitadas (detección de errores pero no recuperación) y posee menores requisitos en cuanto al proceso que implica y al ancho de banda que necesita. Se puede aplicar para servicios de clase B, C o D.

La capa ATM es independiente del medio físico y de los servicios que transporta [I.361]. Se encarga de las funciones relacionadas con la información presente en la cabecera de la celda ATM y que, por tanto, son necesarias para el tratamiento lógico de dicha celda ATM [I.363]. La identificación de canales virtuales y la detección de errores en la cabecera de la celda son ejemplos de dichas funciones.

5.2 ESTRUCTURA DE LA CELDA ATM

La celda ATM tiene 53 octetos de longitud, 5 de los cuales forman la cabecera de la celda que consta de varios campos de información tal como se aprecia en la figura 2.11 [UNI3.0].

La funcionalidad asociada a los distintos campos de la cabecera está reducida al máximo (básicamente encaminamiento) para garantizar un proceso rápido en los nodos de la red. De los cinco octetos que forman la cabecera, tres contienen información de encaminamiento, dos para lo que se denomina identificador de canal virtual (*Virtual Channel Identificator*, VCI) y un octeto para el identificador de camino virtual (*Virtual*

Path Identificator, VPI). El resto de la cabecera consiste en un campo de control de flujo genérico presente sólo en el interfaz entre el usuario y la red (*Generic Flow Control*, GFC), tres bits para un identificador del tipo de información útil que transporta la celda (*Payload Type Identificator*, PTI), un bit de prioridad frente a pérdida de celdas (*Cell Loss Priority*, CLP) y ocho bits para un campo de control de errores en la cabecera (*Header Error Control*, HEC).

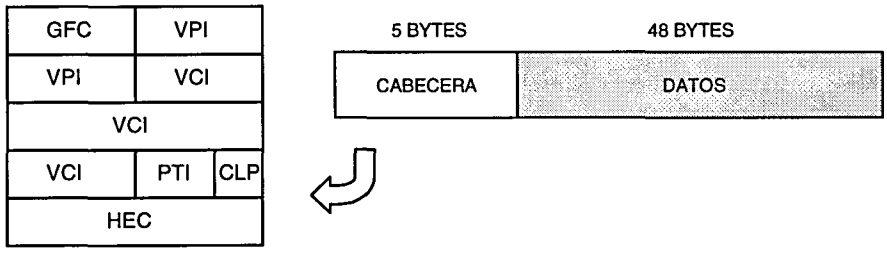

Figura 2.11. Estructura de la celda ATM

Tal y como corresponde a una tecnología de conmutación de paquetes, la cabecera contiene principalmente información para transportar las celdas ATM de un nodo al siguiente. Sin embargo, en lugar de especificarse explícitamente las direcciones de fuente/destino, las celdas ATM se etiquetan mediante números identificadores de conexión virtual (VCI/VPI), lo cual permite identificar las conexiones mediante un número menor de bits y conseguir un encaminamiento universal. El nodo destino se especifica mediante una secuencia de pasos en el proceso de encaminamiento que se determina por procedimientos de señalización en el momento de establecer la conexión.

El desarrollo de una conexión en ATM se divide en tres fases: la fase de establecimiento donde se reservan los recursos necesarios (en el caso en que éstos estén disponibles, en caso contrario la conexión simplemente se rechaza), la fase de transferencia de la información y la fase de liberación de recursos y de la conexión. Las celdas ATM de una misma comunicación viajan por la misma ruta mientras dura la transmisión (ello resulta favorable frente a un modo no orientado a conexión, evitando el resecuenciamiento de paquetes en servicios de tiempo real). Dicha ruta se especifica en la fase de establecimiento de la llamada. La cabecera de la celda ATM contiene en cada momento la información que la red necesita para encaminar la celda sobre la ruta preestablecida.

Un camino virtual consta de varios canales virtuales, como se muestra en la figura 2.12. La distinción en la etiqueta de direccionamiento entre VCI y VPI permite a la red usar una notación más corta y compacta para las rutas con más tráfico (p.e., enlaces entre grandes ciudades) gracias al identificador VPI, a la vez que con el identificador VCI se preserva la identidad de cada canal individual dentro del camino de comunicación. Ello permite que los equipos utilizados en la transmisión traten las llamadas sólo en base al campo VPI, sin la necesidad de operar con el resto del campo

de direccionamiento (VCI) hasta que el enlace general llega a su destino final, donde el tráfico correspondiente a cada canal se distribuye de acuerdo a su VCI.

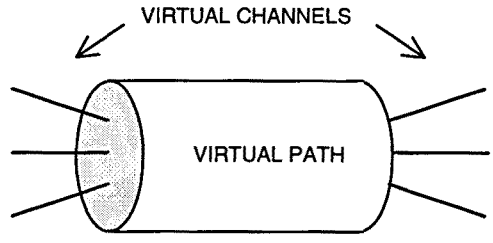

Figura 2.12. Relación entre camino virtual y canal virtual

El resto de campos en la cabecera se usa para controlar el flujo del tráfico generado en la parte de red de usuario (sólo en el interfaz usuario-red), para discernir el tipo de información que transportan las celdas (celdas de usuario o celdas propias de la red), y para marcar cuál es la prioridad de las celdas frente a pérdidas y poder así ejercer alguna acción sobre las mismas en caso de congestión. Finalmente, se utiliza un octeto para asegurar la integridad de la cabecera (básicamente de la dirección) a través de la red.

El ATM usa el concepto de conexión virtual, estableciendo conexiones virtuales entre cada par de nodos de conmutación intermedios en la transmisión desde un extremo fuente a un extremo destino. Estas conexiones se denominan virtuales para distinguirlas de los circuitos o canales dedicados como en el caso del STM. En ATM el enlace no se reserva únicamente a un usuario sino que se comparte temporalmente con otros usuarios.

Las conexiones ATM sólo existen como conjuntos de tablas de encaminamiento que se mantienen en cada conmutador y que están indexadas en función de la etiqueta de direccionamiento (VCI/VPI) de la cabecera de las celdas. Las etiquetas de direccionamiento en ATM tienen sólo un significado local, en cuanto que sólo son relevantes entre dos equipos de conmutación ATM adyacentes, de la misma forma que en tecnologías anteriores como X.25. Cuando se establece un camino virtual, a cada conmutador se le proporciona un conjunto de tablas de búsqueda (indexadas por los identificadores de camino/canal virtual) para identificar una celda de entrada por su etiqueta de direccionamiento presente en la cabecera, encaminarla a través de la red de conmutación hasta su puerto de salida destino, reescribiendo la dirección de entrada por una nueva etiqueta que será usada por el siguiente conmutador en la ruta de transmisión y que de nuevo la reconocerá como un índice de su propia tabla. La celda pasará de conmutador a conmutador por la ruta preestablecida en la fase de establecimiento de la conexión, pero dicha ruta es virtual en cuanto a que las facilidades de transporte sólo se le dedican mientras la celda las atraviese.

5.3 GESTIÓN DEL TRÁFICO EN REDES ATM

La adopción del ATM ha permitido la incorporación de los servicios de velocidad variable en la nueva RDSI-BA. De esta forma se ha conseguido una maximización del uso de los recursos aplicando el principio de ganancia por multiplexación estadística. Como inconveniente, se introducen dos nuevos efectos respecto a las redes de conmutación de circuitos ya existentes: pérdidas de paquetes por desbordamiento de las colas de espera de los elementos de la red, y un retardo variable entre llegadas de paquetes al receptor.

La ubicación de ancho de banda en las redes ATM para servicios CBR es prácticamente igual a la tasa constante de generación. En los servicios VBR se pueden emplear diferentes técnicas de ubicación de ancho de banda [DecFac91]:

- *Ancho de banda de pico.* En esta asignación de ancho de banda se reserva el máximo valor alcanzado por la generación de la tasa binaria, o tasa de pico. Esta asignación es muy ineficiente, ya que, en el transcurso de la conexión, el ancho de banda consumido será muy inferior al reservado. No obstante, esta asignación de recursos asegura una probabilidad de pérdida de información muy baja.
- *Ancho de banda equivalente.* Es una asignación de ancho de banda inferior a la anterior. Depende de las características del servicio y del trayecto que recorrerá la información en la red, de forma que se explotarán las ventajas de la multiplexación estadística al compartir diferentes conexiones los mismos recursos de la red. El ancho de banda equivalente asignado a una conexión será el mayor de los calculados en la compartición de recursos que emplee, a fin de ofrecer un grado de servicio mínimo por parte de la red. El aumento en la eficiencia de la explotación de los recursos de la red, a través de la multiplexación estadística, tiene como contrapartida las pérdidas de información debidas a la capacidad limitada de los buffers de los nodos y las variaciones del retardo extremo a extremo.

Durante el establecimiento de una conexión se estipula un contrato de tráfico entre el servicio y la red. En este contrato se especifican las características del tráfico generado por la fuente y la calidad de servicio ofrecida por la red. A través de los descriptores de tráfico especificados en el contrato usuario-red se determina el encaminamiento de la información y los recursos reservados para la conexión [ReiBer92]. Los descriptores del tráfico empleados en la conexión son [UNI3.0]:

- *Tasa de pico* (*Peak Rate*, R_p). Es el intervalo de emisión mínimo, especificado como el inverso del menor tiempo entre la generación de dos celdas.
- *Tolerancia de la variación del retardo entre celdas* (*Cell Delay Variation Tolerance*, CDVT). Algunos servicios son susceptibles a las variaciones de la

distancia entre celdas resultantes del tránsito en la red. Esta variación se introduce en la espera en las colas de los multiplexores y conmutadores y depende de las tasas de generación instantáneas del conjunto de conexiones que comparten los mismos recursos. En el contrato de tráfico se especifica, a través del parámetro CDVT, el valor máximo de variación admisible por el servicio.

- *Tasa sostenida de celdas* (*Sustainable Cell Rate*, SCR) y tolerancia de ráfaga (*Burst tolerance*, τ_s). Estos parámetros son opcionales y permiten mejorar la explotación de los recursos. La SCR es la expresión de la tasa media generada (R_s) durante un intervalo de tiempo (τ_s). A través de este parámetro se puede controlar la generación sostenida de celdas a tasas inferiores a R_p.

En el contrato establecido entre el servicio y la red se acuerda también la calidad de servicio ofrecida por esta en términos de:

- Probabilidad de pérdida de celdas.
- Retardo máximo extremo a extremo.
- Variabilidad máxima del retardo entre celdas consecutivas (*jitter*).

El dimensionado y ubicación de recursos en redes ATM hace necesario un estudio del comportamiento de los servicios en diferentes escalas temporales, realizando una aproximación jerárquica de tres niveles [CasCav91]:

- *Nivel de llamada.* Se consideran intervalos entre segundos y horas. Este nivel está especialmente indicado para el dimensionado de redes ATM. El dimensionado dependerá de la clase de servicios soportados por la red y la cantidad de éstos que simultáneamente puedan estar en curso. Es imprescindible disponer de las características de estos servicios, relacionadas con el tiempo entre llegadas de nueva petición, la estadística de la duración de la conexión y la dependencia de los puntos de interconexión solicitados en función de la situación geográfica [HofWeb85].
- *Nivel de ráfaga.* Se sitúa en el rango de décimas de segundos a pocos segundos. Tiene especial influencia en el proceso de aceptación de nuevas conexiones por la repercusión en la calidad de los servicios. Este nivel está directamente relacionado con la ubicación de recursos, ya que, dependiendo de la naturaleza de los tráficos que compartan los mismos dispositivos de la red, se alcanzará un grado de servicio ofrecido por la red. El nivel de ráfaga de los diferentes tráficos determinará la probabilidad de que, instantáneamente, se alcancen niveles de utilización de los enlaces superiores a 1 y, por tanto, sea necesario el almacenamiento de las celdas en los buffers de los nodos. El grado de servicio ofrecido dependerá directamente de la capacidad de almacenamiento de los buffers de los dispositivos, estableciéndose una relación de compromiso entre el

tiempo máximo de espera, o retardo admisible, y la probabilidad de pérdida [FuhBou91].

- *Nivel de celda.* Se encuentra en el rango de milisegundos. Determina los valores mínimos en el dimensionado de las colas de espera en los nodos, así como, los rangos de variación del retardo entre celdas [ChaLeo94].

Las funciones de control de tráfico en las redes ATM tienen como propósito la gestión y regulación del tráfico en la red. Estos mecanismos son empleados para el control de congestión. La congestión en una red ATM se alcanza cuando uno de sus recursos es explotado a un nivel próximo a su máximo rendimiento. En general, este nivel de utilización se traduce en niveles de ocupación muy elevados de algunas colas de espera de algunos nodos. El control de congestión debe regular el tráfico entrante en la red, de forma que no se sobrepasen los niveles máximos de ocupación dimensionados. El control de congestión se ejerce de dos maneras diferentes:

- *Control de congestión preventivo.* Las técnicas preventivas de congestión pretenden controlar el tráfico entrante en la red para evitar situaciones de congestión. Para llevar a cabo este control se emplean dos mecanismos:
  - *Control de admisión* (*Call Admission Control*, CAC). Representa el conjunto de acciones tomadas en la fase de negociación del establecimiento de la conexión para aceptar o rehusar una nueva llamada. En base a los recursos disponibles para ubicar el nuevo tráfico, se evalúa la viabilidad de alcanzar la calidad del servicio demandada por la nueva conexión manteniendo las calidades de las conexiones en curso. La aceptación dependerá del contrato usuario-red que se pretenda establecer y del nivel de explotación de los recursos en el momento de la negociación del contrato.
  - *Control de los parámetros de usuario* (*Usage Parameter Control*, UPC). Este control de tráfico pretende asegurar que el tráfico entrante en la red cumple con el contrato establecido en el proceso de aceptación de la llamada. También recibe la denominación de control de policía. Los mecanismos de policía tienen como función monitorizar el tráfico entrante en la red para que se cumpla con el contrato usuario-red establecido en la fase de aceptación de la llamada. La acción del mecanismo de policía se traduce en descartar o marcar las celdas que violen el contrato. La marcación de celdas permite establecer una política de selección cuando un dispositivo está congestionado, a través de la asignación de niveles de prioridad. Así, en estados de congestión, la celdas de prioridad más baja no son almacenadas en las colas de espera.

- *Control de congestión reactivo.* La regulación del tráfico se ejerce cuando la red ha alcanzado un estado de ocupación próximo a la saturación de un recurso. Llegada esta situación, la red puede hacer uso de tres mecanismos:
  - Señalización a las fuentes de tráfico que comparten los recursos congestionados para que disminuyan su tasa de generación.
  - Encaminamiento alternativo de las conexiones sobre otros recursos con un grado de congestión inferior.
  - Descarte de celdas almacenadas en las colas de espera en base al nivel de prioridad y el grado de servicio acordado en la conexión.

6 Redes de acceso

A la hora de proporcionar servicios de vídeo a clientes finales en su residencia particular (*Video To The Home*, VTTH), uno de los aspectos más dificultosos es el problema de la "última milla". Este tipo de servicios, sobre todo si son interactivos, necesita una tecnología de red adecuada para proporcionar la calidad de servicio deseada, y además deben ser proporcionados a bajo coste de forma que sean atractivos a los usuarios. Por supuesto, lo más interesante será ofrecer estos servicios sobre la misma infraestructura que los clásicos de telefonía o de transmisión de datos. Esta infraestructura constituye la red de acceso.

Una de las zonas más problemáticas a la hora de diseñar estándares para estas redes de acceso la constituye Europa. En este continente, la diversidad de estados se refleja en distintas políticas, economías, demografías, etc. Estas diferencias provocan a su vez la existencia de diferentes infraestructuras para las redes de acceso. Si bien la sustitución de la red principal de transporte por sistemas digitales está justificada en términos de coste por usuario, el problema es bien distinto a la hora de llegar hasta las residencias de los usuarios finales. Con motivo de llevar a cabo una estandarización para el futuro de estas redes, el Instituto Europeo de Estandáres de Telecomunicación (*European Telecommunication Standards Institute*, ETSI) ha llevado a cabo diversos trabajos y especificaciones [GilOrt97]. En la actualidad, y teniendo en cuenta que tratamos con servicios multimedia, hablar de nuevas redes de acceso es hablar de redes de acceso de banda ancha. Los elementos de una red de acceso de banda ancha se presentan en la figura 2.13 [Fis98].

Distintos tipos de acceso han sido ya especificados, con distintas características entre ellos que los hacen adecuados en unos entornos mientras que en otros son menos válidos. La mayoría de arquitecturas comparten los mismos elementos principales. En [ATMFORUM1] se presentan varios ejemplos. El nodo de acceso (*Access Node*, AN) realiza la conexión con la red principal de transporte, generalmente basada en ATM, realizando funciones de conversión de protocolos y velocidades de transmisión. La red

de distribución se encarga de transmitir las señales hasta los usuarios finales. Finalmente, el elemento de terminación de red (*Network Termination*, NT) es la frontera con el usuario privado, pudiendo ser pasivo o activo. Para gran parte de los servicios, el PC multimedia sería el terminal idóneo a escoger. Sin embargo, se vaticina que para servicios de difusión de vídeo o de VoD el terminal más popular será el clásico aparato receptor de TV junto con un adaptador (*Set Top Box*, STB).

Dentro de las redes de difusión existen multitud de posibilidades en función de la combinación de tecnologías utilizadas para el alimentador (*feeder*) y para la transmisión en la última milla. Tan sólo algunas de ellas tendrán su lugar finalmente en el mercado. En los siguientes apartados se introducen algunas de estas tecnologías.

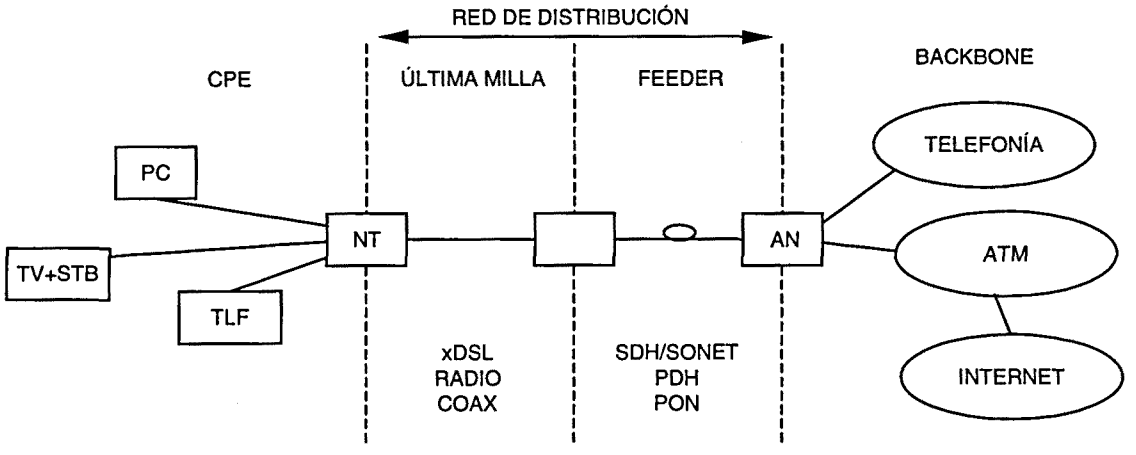

Figura 2.13. Modelo genérico de red de acceso de banda ancha

6.1 ESCENARIO ADSL/HDSL/SDSL: FTTE

El escenario FTTE (*Fibre To The Exchange*) ofrece acceso de banda ancha a través de los pares trenzados de cobre instalados entre la central y el abonado. En el alimentador, conectado a la red principal de transporte mediante un conmutador o router de acceso, se utiliza fibra y jerarquía digital síncrona (*Synchronous Digital Hierarchy*, SDH) o plesiócrona (*Plesiochronous Digital Hierarchy*, PDH). La interconexión entre el alimentador y los bucles de abonado se llevan a cabo mediante un multiplexor (*Digital Subscriber Line Access Multiplexor*, DSLAM).

La línea digital de abonado asimétrica (*Asymmetric Digital Subscriber Line*, ADSL) puede cubrir muy bien las necesidades de usuarios particulares, al no ser necesarias altas tasas de bit en el sentido cliente-red. Por otra parte, las empresas y compañías se pueden decantar más por una tecnología simétrica (*Symmetric DSL*, SDSL) o de alta velocidad (*High-Speed DSL*, HDSL). En la tabla 2.8 se esquematizan los principales parámetros de estas tecnologías.

Teconología	Canal de subida	Canal de bajada	Alcance	Modulación
ADSL	hasta 640 Kbps	hasta 8 Mbps	hasta 5 Km.	DMT/CAP
ADSL light / UDSL	hasta 176 Kbps	hasta 1.5 Mbps	hasta 5 Km.	DMT/CAP
SDSL	hasta 1.5 Mbps	hasta 1.5 Mbps	hasta 4 Km.	CAP/DMT
VDSL	hasta 20 Mbps	hasta 52 Mbps	hasta 1.5 Km.	QAM/DMT
HDSL	hasta 2 Mbps	hasta 2 Mbps	hasta 3.5 Km.	2B1Q

Tabla 2.8. Tecnologías xDSL

6.2 ESCENARIO VDSL/PON: FTTC

En este escenario, fibra hasta el distribuidor (*Fibre To The Curb*, FTTC) se plantea el uso de tecnología de muy alta velocidad en la línea de abonado (*Very high-speed* DSL, VDSL) y de una red óptica pasiva (*Passive Optical Network*, PON) en el alimentador. De esta forman se pueden ofrecer hasta 52 Mbits de bajada al cliente, pero con alcances más reducidos.

6.3 ESCENARIO FTTH

En este escenario se plantea la posibilidad de hacer llegar fibra hasta los domicilios (*Fibre To The Home*, FTTH). De esta forma se podrían llevar hasta el usuario final tasas de bit extremadamente altas.

Los altos costes iniciales de implementación se podrían ver a largo plazo compensados por los costes más bajos de mantenimiento. Por otra parte, las funciones del NT y la unidad de red óptica (Optinal Network Unit, ONU) se realizarían en un solo dispositivo.

6.4 ESCENARIO HÍBRIDO: HFC

La solución híbrida (*Hybrid Fibre Coax*, HFC) aparece como la mejor para las empresas proveedoras de TV por cable con el objetivo de suministrar servicios de banda ancha interactivos. La infraestructura existente de cable quedaría sin tocar y se utilizaría para transmisión bidireccional.

En este caso se haría necesario un NT a un extremo del alimentador y de un ONU al otro extremo. Para la interconexión entre la red principal de transporte y el alimentador se utilizará un elemento terminal de línea óptica (*Optical Line Terminator*, OLT).

6.5 ESCENARIO RITL

Otra posibilidad consiste en el establecimiento de un enlace radio para dar el servicio de última milla (Radio In The Loop, RITL).

Una estación base da servicio a un número determinado de clientes dentro de una celda. En este caso no está contemplada la posibilidad de la movilidad de los clientes.

6.6 ESCENARIO PLC

Como última posibilidad para la tecnología de la última milla comentaremos el escenario en el cual se hace uso de la línea de transmisión de energía eléctrica para la comunicación interactiva con las residencias de los clientes (*Power Line Communication*, PLC). Estas líneas se pueden utilizar en la última milla para alcanzar la residencia destino, así como dentro de la propia residencia.

7 Transmisión de vídeo sobre redes de banda ancha

Una vez analizados los conceptos más importantes de las técnicas de compresión de vídeo, de la B_ISDN y de las redes de acceso, en este apartado se analizan algunos métodos de adaptación entre ambos. Para ello se introducirán algunos modelos propuestos para la transmisión de vídeo sobre redes de banda ancha. Dichos modelos serán los siguientes:

- Especificación de Vídeo bajo Demanda del ATM Forum
- Recomendación J.82 de la ITU-T
- Especificación DAVIC

7.1 ESPECIFICACIÓN DE VÍDEO BAJO DEMANDA DEL ATM FORUM

7.1.1 Definición y configuración

Esta especificación trata sobre el acuerdo de implementación para el transporte de audio, vídeo y datos codificados en MPEG-2 sobre redes ATM como soporte de servicios multimedia audiovisuales (*Audio-visual Multimedia Services*, AMS), realizado por el ATM Forum. La última especificación data de marzo de 1997, llevada a cabo por el grupo de trabajo técnico "Aspectos de Servicio y Aplicaciones".

El servicio VoD es un servicio asimétrico que supone varias conexiones. Proporciona la transferencia de información de vídeo codificado y comprimido desde un servidor hasta un cliente. En el receptor, la información es reensamblada, descomprimida, decodificada, pasada a forma analógica y visualizada en un monitor. Se trata de un servicio interactivo, en el cual el usuario tiene cierto control sobre el material que desea ver así como el momento de la visualización. Está destinado principalmente a funciones de entretenimiento permitiendo al usuario conectarse a una base de datos donde existe una librería de secuencias (p.e., películas, documentales, ...). Las conexiones se establecen mediante señalización entre el usuario y la red. La transmisión de programas se hace fundamentalmente punto a punto entre el proveedor de información (*Video Information Provider*, VIP) y el usuario final. Otras características añadidas pueden permitir al usuario las funciones clásicas de un reproductor de vídeo doméstico clásico, como pausa, rebobinado, avance rápido, etc.

Para que un servicio sea considerado auténtico VoD se deben cumplir las tres siguientes premisas [LiaLi97]:

- Posibilidad de ver cualquier vídeo de los disponibles en el servidor.
- Posibilidad de hacer uso del servicio en cualquier momento del día.
- Posibilidad de realizar todas las operaciones características de un VCR: pausa, avance rápido,...

En el caso de que alguna de las premisas anteriores no se cumpla, el servicio se conoce como vídeo casi bajo demanda (Near Video on Demand, NVoD).

Al margen de la especificación del ATM Forum, son numerosos los trabajos de investigación proponiendo nuevas mejoras y soluciones para este tipo de servicio. Por ejemplo, en [LiaLi97] se comenta la posibilidad de reducir el coste del sistema por usuario mediante el agrupamiento de usuarios, si bien esta modalidad llevaría al servicio NVoD. Para evitar esta degradación, se presenta un nuevo protocolo denominado *Split and Merge* (SAM). Otro interesante trabajo, a modo de tutorial, se centra en la disposición de servidores en paralelo para montar sistemas a gran escala dando servicio a un gran número de usuarios [Lee98].

La fase 1 de la especificación del ATM Forum está dirigida especialmente al servicio VoD mediante MPEG-2 SPTS con tasa de paquetes constante (*Constant Packet Rate*, CPR), especificando [ATMFor1]:

- Requisitos AAL
- Encapsulado de MPEG-2 TS en PDUs del AAL-5
- Señalización ATM y control de las conexiones
- Características del tráfico
- Calidad de servicio

La configuración de servicio propuesta a título informativo es la esquematizada en la figura 2.14.

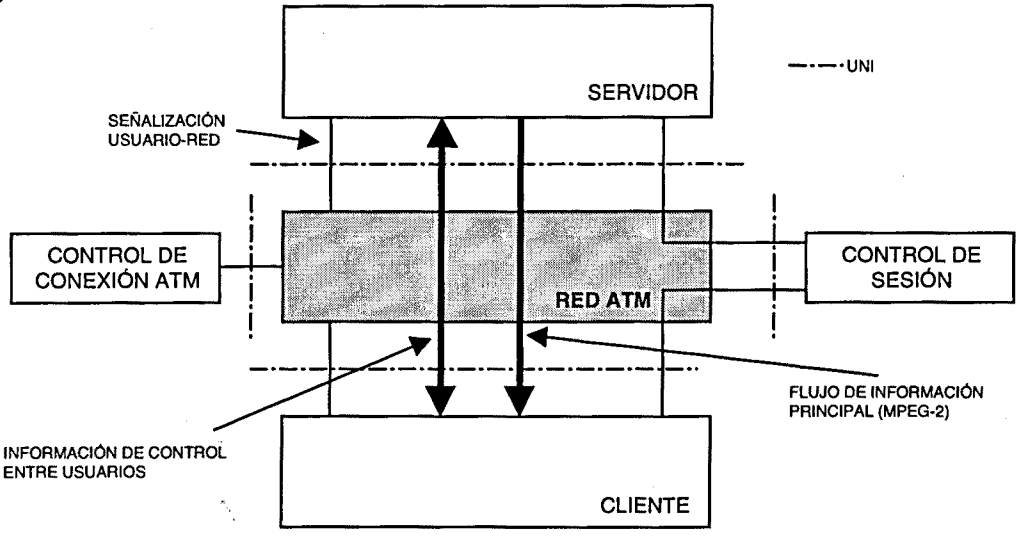

Figura 2.14. Configuración de referencia VoD

La red ATM puede estar basada en varias tecnologías (HFC, FTTC, FTTH, etc) y presentar diversas topologías (algunos ejemplos se pueden consultar en el apéndice B de [ATMFor1]).

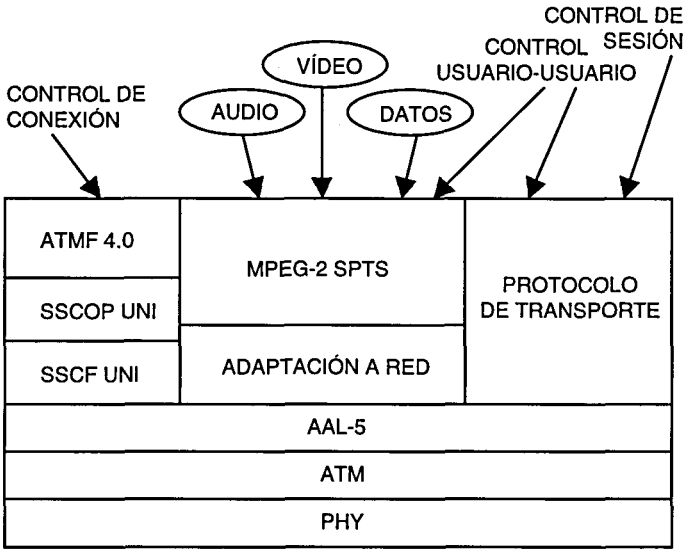

Figura 2.15. Modelo de referencia de protocolos ATM Forum VoD

En la figura 2.15 se presenta el modelo de referencia de protocolos para el ATM Forum VoD. Uno de los aspectos más importantes del acuerdo de implementación VoD es la definición de cómo se transmiten los paquetes MPEG-2 TS por la red. Se define el nivel de adaptación ATM 5 con un subnivel de convergencia específico de servicio nulo para el transporte de los paquetes MPEG-2. Por defecto, se envían dos paquetes de transporte MPEG-2 dentro del campo de datos de la PDU del AAL-5, cuando la red está utilizando conexiones virtuales permantentes (*Permanent Virtual Connections*, PVC). Utilizando conexiones virtuales conmutadas (*Switched Virtual Connections*, SVC), el valor de este número de paquetes es negociable durante la fase de establecimiento del circuito virtual, utilizando la señalización 4.0 del ATM Forum [ATMFor2]. Como nivel básico de interoperabilidad, se define que todos los equipos deben soportar el valor por defecto de dos paquetes por SDU AAL-5. Esta situación queda reflejada en la figura 2.16.

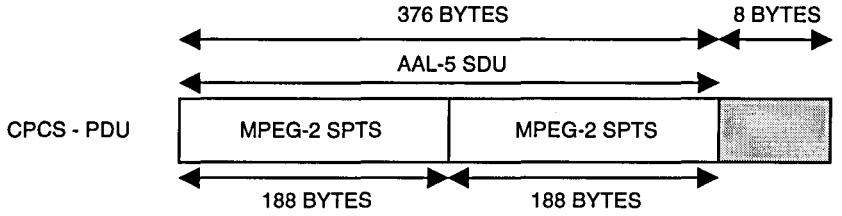

Figura 2.16. Encapsulado de 2 paquetes MPEG-2 en una SDU AAL-5

Cuando se recibe una PDU AAL-5 errónea pero de longitud correcta, las prestaciones del servicio se pueden mejorar si en lugar de descartarla se entrega al siguiente nivel con una indicación del error. Esta situación puede ser tratada más

eficazmente por el decodificador MPEG-2 que si no recibiese ningún dato. Experimentos que corroboran estas afirmaciones se han realizado en [GriKha98]. En el caso de que la longitud de la PDU sea errónea, es preferible descartarla completamente, o bien utilizar alguna técnica de relleno.

7.1.2 Parámetros de tráfico y calidad de servicio

En cuanto a los parámetros de tráfico, la fuente MPEG-2 se considera un flujo de información CPR. El tráfico a la salida del codificador debe ser conformado para cumplir con el contrato de tráfico CBR negociado con la red. Si la tasa de paquetes del MPEG-2 SPTS es de *M* por segundo, utilizando la especificación por defecto de 2 paquetes por SDU AAL-5, se tiene:

$$\textit{Tasa de pico nivel ATM} = 4M \text{ celdas/segundo} \quad (2.2)$$

Además, el jitter introducido deberá cumplir el contrato negociado con la red en términos CDVT. En caso contrario, las celdas serían descartadas. Las principales causas introductoras de CDV son:

- Multiplexación de diferentes CV sobre un mismo UNI: una simple aproximación nos permite afirmar que, para un servidor de vídeo concreto, el jitter máximo introducido debido a multiplexación se producirá cuando todos los canales virtuales destinados a un mismo UNI lleguen a la función de multiplexación al mismo tiempo. Si hay un número n de dichos canales virtuales, este fenómeno supondrá un CDV de n-1 slots.
- Overhead de nivel físico: Aproximable por 1 slot, o incluso menos.
- Soporte a celdas OAM: También aproximable por 1 slot.

Así, para un UNI transportando *n* VCs, se debe adoptar un valor *n+1* para la CDVT.

En cuanto a la calidad de servicio, se deben especificar dos parámetros de retardo:

- CDV pico a pico
- CTD máximo

Además, se deben especificar los tres siguientes parámetros:

- Tasa de pérdidas de celdas (Cell Loss Ratio, CLR)
- Tasa de error de celdas (Cell Error Ratio, CER)
- Tasa de bloques de celdas muy erróneos (Severely Errored Cell Block Ratio, SECBR)

7.2 RECOMENDACIÓN ITU-T J.82

La ITU-T ha realizado también su propia recomendación para el encapsulado de secuencias MPEG en celdas ATM [I.J82]. Por una parte describe un método utilizando AAL-5 como el descrito anteriormente para el ATM Forum VoD. Por otro lado,

propone un método alternativo utilizando el nivel de adaptación 1. La técnica es algo más compleja, pero ofrece mayor protección frente a los problemas que pueda causar la red ATM. Para ello, se proponen ciertos esquemas de corrección de errores.

Por una parte, la utilización del AAL-1 permite enfrentarse al problema del jitter de una manera más eficiente, al utilizar marcas temporales. Con estas marcas se puede simular un retardo constante extremo a extremo. Además, este nivel de adaptación incluye números de secuencia, lo que permite detectar pérdidas de celdas y celdas mal insertadas.

Para el control de errores se propone la utilización de los códigos de Reed-Solomon entrelazados. El código utilizado utiliza 4 bytes de redundancia sobre 124 de información, como se muestra en la figura 2.17, donde se muestra también el entrelazado propuesto.

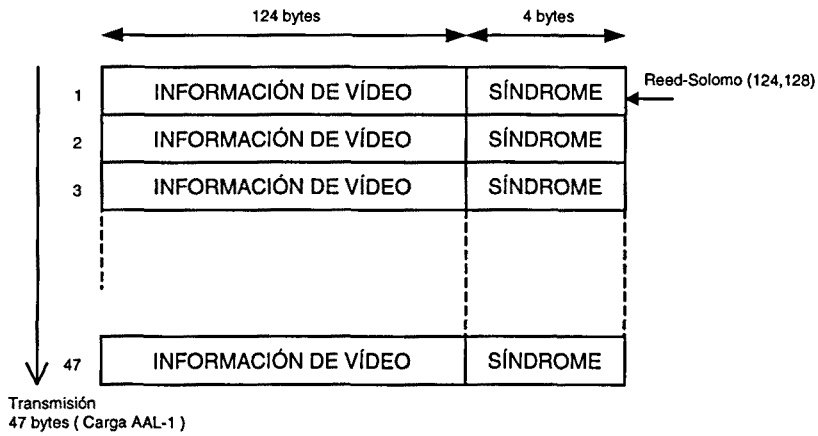

Figura 2.17. Aplicación de código Reed-Solomon y entrelazado

7.3 DIGITAL AUDIO VISUAL COUNCIL (DAVIC)

DAVIC es una asociación sin ánimo de lucro con sede en Suiza, formada por alrededor de 175 compañías procedentes de más de 25 paises. Representa todos los sectores de la industria audiovisual, desde fabricantes hasta proveedores de servicios. Por citar algunos de los miembros, se puede comentar la presencia de Philips, Sun Microsystems, HP, IBM, Sony, AT&T, Bell Atlantic, BBC, Microsoft, Oracle, France Télécom, Deutsche Telekom, etc. Se creó en 1994, con el objetivo de promover el éxito de los servicios y aplicaciones audiovisuales digitales interactivos. Para ello se proponen especificaciones de interfaces abiertos y de protocolos que maximicen la interoperabilidad a través tanto de límites geográficos como a través de diversas aplicaciones, servicios e industrias [DAVIC1]. Básicamente, su trabajo consiste en identificar, seleccionar y desarrollar especificaciones de interfaces, protocolos y arquitecturas de servicios y aplicaciones audiovisuales digitales. Para ello, se busca también el reconocimiento y aprobación por parte de las organizaciones clásicas de estándares. Así, las fuentes para los trabajos de DAVIC provienen del ATM Forum, ITU, ISO/IEC, MPEG, ANSI, etc.

Entre las prioridades actuales de DAVIC se encuentran las siguientes [Tho98]:

- Apuntalamiento del éxito de MPEG como primer estándar global de TV.
- Apuesta por el equipo receptor de TV y *set top box* en lugar del PC multimedia.
- Promover la evolución de redes digitales residenciales y sistemas para entrega basada en IP y almacenamiento local.

En la actualidad, son ya bastantes los ejemplos de aplicaciones y servicios que utilizan las especificaciones DAVIC alrededor del mundo (Panasonic Multimedia on Demand System, Philips' Clevercast™, IBM's Logicast™, ...).

7.3.1 Modelo de referencia

El tipo de sistema al que están dirigidas las especificaciones DAVIC se compone de cinco entidades [DAVIC2], según se observa en la figura 2.18.

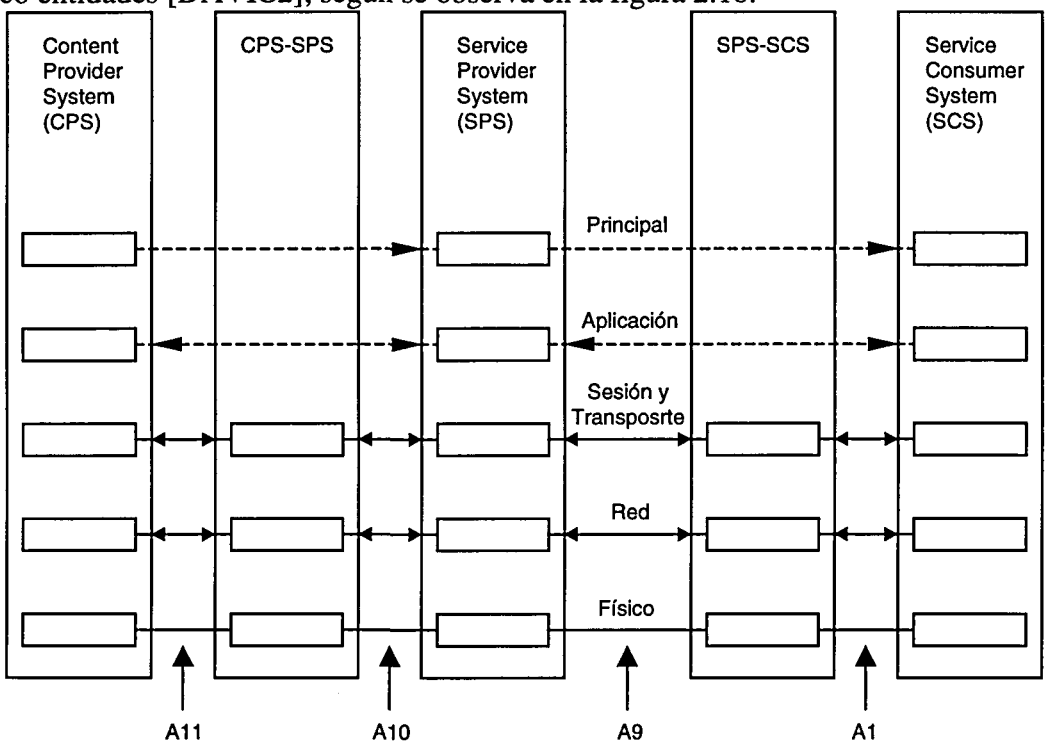

Figura 2.18. El sistema general DAVIC

Un ejemplo práctico podría consistir en un estudio de TV suministrando programas a una compañía de TV que está ofreciendo un servicio de VoD a sus clientes. El estudio de TV sería el CPS, la compañía de TV el SPS y por supuesto el usuario final sería el SCS, mediante su *set top box*. Los sistemas de entrega están constituidos básicamente por las redes de interconexión. En el ejemplo anterior, el sistema de entrega entre el proveedor del servicio (compañía de TV) y el usuario final podría ser un acceso HFC, mientras que la comunicación entre el estudio y la compañía podría ser mediante una red ATM basada en fibra óptica.

El modelo DAVIC proporciona interfaces bien definidos, llamados puntos de referencia y numerados desde A0 hasta A11. En la tabla 2.9 se describen los interfaces mostrados en la figura 2.18.

Punto de Referencia	Descripción
A1	Interface entre el consumidor del servicio y el sistema de entrega SPS-SCS. Utilizado para la conexión de la unidad set top al acceso de red (p.e., modem ADSL)
A9	Interface entre el sistema proveedor de servicio y el sistema de entrega SPS-SCS. Para un servidor de VoD podría ser un enlace ATM basado en SDH/SONET
A10	Interface entre el sistema proveedor de servicio y el sistema de entrega CPS-SPS
A11	Interface entre el sistema proveedor de contenido y el sistema de entrega CPS-SPS

Tabla 2.9. Puntos de referencia DAVIC

7.3.2 Flujos de información DAVIC

Se definen cinco flujos de información, de los cuales uno se encarga de transportar el contenido solicitado por el usuario y los otros cuatro del control y gestión. Estos cinco flujos, numerados de S1 a S5, se describen en la tabla 2.10.

Flujo DAVIC	Definición
S1	Flujo de contenido
S2	Control de aplicación
S3	Control de sesión
S4	Control de enlace
S5	Flujo de gestión

Tabla 2.10. Flujos de información DAVIC

El flujo S1 transporta el contenido del servicio, como por ejemplo una película codificada MPEG. Una posibilidad, dependiendo de la tecnología de red, sería transmitir la pila de protocolos MPEG sobre ATM, como se muestra en la figura 2.19.

El flujo S2 está relacionado con el S1, pero se utiliza como canal de control para el contenido del servicio. Por ejemplo, en un servicio VoD, la información referente a solicitudes de avance rápido o pausa viajaría en este flujo.

El flujo de control S3 se utiliza para configurar y controlar las sesiones de aplicación entre un servidor y un cliente. Siguiendo con el ejemplo anterior, S3 se habría utilizado para crear las conexiones S1 y S2 entre el cliente y el servidor.

El control de conexión a nivel de red entre cliente y servidor se realiza en el flujo S4. Se utiliza para establecer y liberar conexiones entre el servidor (o cliente) y el sistema de entrega. Finalmente, el flujo S5 se destina a la gestión de red y sistema, utlizando por ejemplo el protocolo SNMP (*Simple Network Managenent Protocol*).

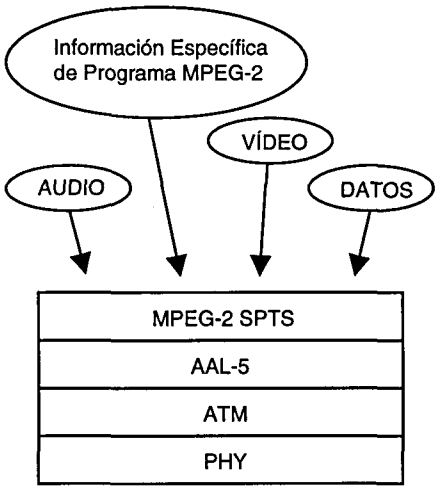

Figura 2.19. Encapsulado S1

7.3.3 Servicios multimedia

Aparte de las especificaciones sobre funcionalidad y requisitos de las aplicaciones comunes, DAVIC describe con detalle ejemplos para las aplicaciones multimedia más comunes, como:

- Películas bajo demanda (*Movies on Demand*, MoD)
- Teletienda (*Teleshopping*)
- Difusión (*Broadcast*)
- Vídeo casi bajo demanda (*Near Video on Demand*, NVoD)
- Difusión a posteriori (*Delayed Broadcast*)
- Juegos (*Games*)
- Trabajo a distancia (*Telework*)
- ...

Para cada uno de estos servicios, se proporciona una descripción detallada, una serie de especificaciones base, incluyendo las funciones de cada una de las entidades que intervienen en el servicio (contenedor de la información, proveedor del servicio, usuario final, ...), características y extensiones.

CAPÍTULO 3

CARACTERIZACIÓN DEL TRÁFICO DE VÍDEO MPEG VBR COMO PROCESO AUTOSEMEJANTE

Recientes medidas de tráfico de vídeo VBR, así como del tráfico generado en redes de área local, canales de control de la red digital de servicios integrados, y otros sistemas de comunicación, revelan una naturaleza autosemejante en su comportamiento. Esta propiedad se manifiesta de varias formas. Por una parte, es posible observar como las ráfagas de estos tipos de tráfico se comportan de la misma forma en un amplio margen de escalas de tiempo. Además, las funciones de autocorrelación de los procesos definidos por estos tráficos decaen de una forma más lenta que una función exponencial, indicando una fuerte dependencia a largo plazo. Finalmente, la suma del valor de la función de autocorrelación para todos los desplazamientos posibles no converge a un valor finito, lo cual, en términos frecuenciales, implica la divergencia del valor de la función de densidad espectral de potencia en el origen.

Por otra parte, se ha comprobado que la presencia de autosemejanza en el tráfico a transmitir ha de ser tenida en cuenta en el dimensionado de los dispositivos de red. En general, los modelos clásicos de tráfico no capturaban correctamente las dependencias a largo plazo. Este hecho lleva frecuentemente ha resultados demasiado optimistas cuando se trata de dimensionar las longitudes de los buffers de almacenamiento de multiplexores y conmutadores y las capacidades de los enlaces que los unen.

1 Introducción

El comportamiento del tráfico de vídeo MPEG VBR es objeto de estudio en multitud de trabajos recientes. Entre sus propiedades más importantes aparece un nuevo fenómeno, con implicaciones directas en diversos campos, como por ejemplo en el modelado de dicho tráfico, o en el diseño y control de redes de banda ancha. Este fenómeno, que será ampliamente tratado en este capítulo, consiste en la similar apariencia del tráfico al ser observado en distintas escalas de tiempo. Básicamente, se observa como el comportamiento a ráfagas del tráfico se repite sobre un amplio rango de dichas escalas temporales. En otras palabras, cuando se va promediando el tráfico sobre intervalos cada vez mayores, el aspecto que presentan las medidas realizadas continúa siendo a ráfagas, en lugar de tender a una secuencia plana. Dicha naturaleza *fractal o autosemejante* aparece también en otros tipos de tráfico, destacando el tráfico generado por redes de área local (*Local Area Networks*, LAN), el tráfico en los canales de control de la ISDN, o el tráfico procedente de los servidores de información en Internet (*World Wide Web*, WWW). Un parámetro clave en el entorno de los procesos autosemejantes es el llamado *parámetro de Hurst (H)* [TsyGeo97], el cual está definido para capturar el grado de autosemejanza de un proceso.

De entre los estudios comentados, debe destacarse por una parte el análisis de cientos de millones de tramas en diferentes LANs Ethernet en el Bellcore Morristown Research and Engineering Center [LelTaq93][LelTaq94]. Este trabajo se inicia con una detallada revisión teórica de las propiedades de los procesos autosemejantes. Posteriormente presenta el comportamiento autosemejante del tráfico de las redes Ethernet. Para ello, las medidas efectuadas se hicieron a lo largo de varios años, cambiando tanto la topología de la red como los puntos de medida. Por otra parte, el tráfico en redes de área extensa (*Wide Area Networks*, WAN), ha sido analizado en [PaxFlo95], mediante el estudio de trazas de tráfico recogidas en Berkeley. Dicho estudio permite afirmar que el tiempo entre los establecimientos de nuevas conexiones (TELNET, FTP, ...) queda bien modelado mediante una distribución exponencial. Sin embargo, las llegadas de nuevos paquetes dentro de las conexiones se alejan de ese modelo. Además, multiplexar las llegadas de varias conexiones empeora aún más las cosas, con lo que de nuevo se llega a la necesidad de la utilización de modelos autosemejantes. También son destacables los estudios realizados en [MeiWir91] y en [CroBes95] para tráfico ISDN y World Wide Web (WWW) respectivamente. En este último trabajo se muestra como el tráfico provocado por el acceso de usuarios a servidores web se comporta también de forma autosemejante cuando la carga es alta. Para ello, analizan la distribución de los tiempos de actividad e inactividad,

comprobando que en ambos casos es del estilo "abierto" (*heavy tailed*). Finalmente, en cuanto al tráfico de vídeo VBR, son numerosos los trabajos que exponen la presencia, junto a las clásicas dependencias a corto plazo (*Short Range Dependence*, SRD), de dependencias a largo plazo (*Long Range Dependence*, LRD) en su función de autocorrelación, lo cual es otra manifestación de comportamiento autosemejante. Intuitivamente, un proceso con LRD exhibe una memoria larga, es decir, la dependencia entre eventos que están separados disminuye muy lentamente al aumentar la distancia. Entre estos estudios se encuentran los excelentes trabajos de Mark Garret y otros en [Gar93][GarWil94], así como las medidas realizadas en [BerShe95].

Así, los dos efectos fundamentales que definen estos procesos autosemejantes son los siguientes:

- *Efecto Noah:* Funciones de densidad de probabilidad abiertas de algunas variables, como por ejemplo la longitud de las ráfagas de tráfico o el tiempo entre llegadas.
- *Efecto Joseph:* Dependencia a largo plazo en funciones de autocorrelación.

En la tabla 3.1 se presentan algunas de las propiedades de los procesos fractales en comparación con los procesos convencionales [Wan98].

	Procesos convencionales	Procesos fractales
Escalas de tiempo	Limitadas	Amplio rango
Estadísticas	Bien definidas	Regularmente definidas
Dependencias	Corto plazo	Largo plazo
Decaimiento de varianzas	Rápido	Lento
Espectro de potencia en origen	Positivo y finito	Divergente

Tabla 3.1. Características de procesos convencionales y fractales

Diferentes estudios muestran, por otro lado, que la presencia de tráfico autosemejante en las colas de los multiplexores y conmutadores de las redes de banda ancha implica la necesidad de un mayor número de recursos para mantener la misma calidad de servicio (Quality of Service, QoS) que cuando el tráfico no es autosemejante. La LRD implica que puede darse una persistencia considerable en procesos de tráfico a ráfagas. El crecimiento de los elementos almacenados en las colas de espera de los dispositivos de las redes de conmutación de paquetes se debe a que el tráfico incidente es en ocasiones superior al que puede ser servido. En ausencia de LRD, es muy improbable que esta situación se extienda en un periodo largo de tiempo. Sin embargo, la presencia de LRD provoca que los episodios de congestión puedan extenderse, aumentando la probabilidad de pérdida o el retardo introducido por estos dispositivos. En [AdaMuk95] y [ConGre96] se obtienen estas conclusiones tras la realización de diferentes simulaciones. Por otra parte, en [TsyGeo97] se proporcionan cotas para la probabilidad de pérdida. Además, se lleva a cabo una revisión de las definiciones

clásicas de autosemejanza, demostrando la existencia de conceptos redundantes, y proporcionando una nueva y más compacta definición del fenómeno autosemejante.

Los modelos clásicos para el tráfico en redes de paquetes no capturan el fenómeno autosemejante. Esto lleva a errores cuando se utilizan dichos modelos para conducir simulaciones con el fin de dimensionar el tamaño de los buffers y la capacidad de los enlaces de las redes de paquetes. Por lo general, estos modelos llevan a resultados demasiado optimistas, es decir, los tamaños obtenidos para los buffers son menores de los que realmente se van a necesitar para una QoS determinada. Por tanto, se hacen necesarios nuevos modelos que capturen el comportamiento autosemejante. En [AdaMuk95] las simulaciones son conducidas por un modelo autoregresivo, integrativo y de media móvil (AutoRegresive, Integrative and Moving Average, ARIMA) fraccional, anulando la componente de media móvil y tomando orden 1 para la componente autoregresiva. Este modelo genera un proceso autosemejante, pero no es válido como modelo para el tráfico de vídeo VBR MPEG ya que no captura adecuadamente la dependencia a corto plazo ni el patrón que se repite debido a los modos de codificación del algoritmo MPEG. Modelos que se ajustan bien al tráfico de vídeo MPEG VBR se han presentado en [CasLor97] para el nivel de escena, y en [Mat96] y [CruFer98] para el nivel de grupo de cuadros (Group of Pictures, GoP). En [ConGre96] se trabaja con una secuencia de tráfico real ("*Star Wars*") y con un modelo basado en cadenas de Markov que es necesario ajustar según el tipo de secuencia de vídeo que se desee modelar.

Por otra parte, en [Hos84] y [LauErr95] se presentan los dos algoritmos más clásicos para la generación de procesos autosemejantes. Ambos serán presentados posteriormente dentro de este mismo capítulo.

El resto de este capítulo está organizado de la forma siguiente. En primer lugar se hace un pequeño repaso sobre el modelado de tráfico, para continuar con el estudio de las características de los procesos autosemejantes y sus posibles definiciones. Posteriormente se introducen los métodos clásicos de generación de procesos con LRD: el algoritmo de Hoskings y el algoritmo RMD. Estos métodos de generación presentan algunos inconvenientes, como el elevado tiempo de computación necesario o la gran cantidad de memoria requerida. En este trabajo se presenta un nuevo método que permite solucionar estos inconvenientes, con lo cual puede ser utilizado para generar tráfico utilizado en simulaciones en tiempo real. La validez del método se comprobará analítica y experimentalmente. A continuación, se utilizará el método propuesto para el ajuste de un nuevo modelo ARIMA fraccional para el tráfico de vídeo MPEG VBR a nivel de cuadro.

2 Modelado de tráfico

La confección de modelos de tráfico de los servicios soportados por las redes ATM es un objetivo necesario para el dimensionado de los componentes de estas redes y para la evaluación de las prestaciones de los dispositivos que las componen. En particular, el modelado del tráfico de vídeo digital codificado es especialmente importante, ya que los servicios de vídeo bajo petición y distribución de vídeo, además de tener una demanda en continuo crecimiento, son los mayores consumidores de ancho de banda de la red.

A través de los modelos de tráfico se pueden hallar unos descriptores de tráfico adecuados que caractericen un servicio, con lo que se facilitan las labores de gestión de redes ATM. La aportación de modelos de fuentes de tráfico permite, entre otros:

- Dimensionar la red para soportar una carga de tráficos heterogéneos simultáneos.
- Evaluar las prestaciones de los dispositivos o del comportamiento de la red extremo a extremo.
- Establecer criterios de control de admisión de nuevas llamadas con un nivel de calidad de servicio especificado.
- Determinar un control de congestión preventivo que monitorice el comportamiento de la conexión de forma que se respete el contrato usuario-red.
- Definir las funciones reguladoras del control reactivo y los umbrales de actuación de éste.
- Predecir el comportamiento del tráfico, simple o multiplexado, para aumentar el grado de servicio ofrecido y la explotación de los recursos.

Un modelo de tráfico debe capturar los comportamientos del tráfico generado por el servicio que son significativos a la hora de desarrollar las funciones especificadas anteriormente. La bondad del ajuste de un modelo debe ser evaluada en tanto en cuanto capture estos comportamientos. La gran mayoría de los modelos propuestos intentan caracterizar el comportamiento de uno o varios de los parámetros relacionados con la tasa de generación (λ) [And93].

Los modelos pueden ajustar distintos parámetros para capturar el comportamiento del tráfico generado por los servicios en diferentes niveles temporales. Cabe distinguir tres niveles temporales [GihGia91]:

- Nivel de llamada o duración de la conexión.
- Nivel de ráfaga o variación de la actividad de la conexión.
- Nivel de celda relacionado con el tiempo entre llegadas de celdas.

El nivel de llamada ha sido caracterizado, en general, por procesos de Markov y se ha comprobado que este modelo es válido para el dimensionado de redes [Hui90]. El

nivel de ráfaga es el más analizado por su impacto en la ubicación de recursos y la calidad de servicio y, por tanto, su caracterización es fundamental. El nivel de celda es considerado en algunos estudios, con el fin de establecer un análisis detallado del comportamiento de los dispositivos. El tamaño de las colas de almacenamiento en los nodos y multiplexores reduce sustancialmente el interés del estudio de este nivel en la ubicación de recursos. El estudio del nivel de celda contribuye al análisis de la variación del retardo entre celdas consecutivas, al diseño de estrategias de sincronización (p.e., los parámetros de los PLL digitales) y al dimensionado de los buffers de contención de multiplexores y conmutadores.

Los diferentes modelos propuestos en la literatura se pueden clasificar, o bien por los parámetros de tráfico que ajustan, o bien por el nivel temporal donde son aplicados. A su vez, admiten clasificaciones en modelos continuos o discretos, atendiendo al tipo de síntesis realizada.

Las series temporales generadas por los modelos de tráfico son eventos que pretenden definir la tasa instantánea de generación o la tasa media de generación en un intervalo dado. Los procesos de generación de tasa media en intervalos de duración dados proporcionan como eventos el volumen de información a transferir en un intervalo, mientras que los procesos de generación de llegadas hacen hincapié en como se producen las trasferencias de información indicando el tiempo entre dos llegadas consecutivas de paquetes de información. En el caso de modelos en tiempo discreto los valores generados son números enteros positivos. Los modelos en tiempo continuo operan con números reales, aunque, posteriormente pueden ser truncados cuando se emplean en simulaciones. Dentro de los trabajos presentados en la literatura se han desarrollado también modelos compuestos. Estos modelos conjugan la generación de tasas en intervalos dados y tasas instantáneas. Se basan en desarrollar un proceso que sintetice el tiempo entre llegadas y otro proceso que determine el número de llegadas en ese instante. Se puede denominar procesos de llegada en grupo (Batch Arrival Processes, BAP).

2.1 MODELOS ARIMA (p,d,q)

Estos procesos han sido ampliamente estudiados en la literatura y en su forma más general se denominan procesos autoregresivos, integrativos de media móvil (*autoregresive integrative moving average*, ARIMA) [BoxJen76]. Los modelos autoregresivos se emplean en el contexto de fuentes de tráfico sintéticas o en predicción de tráfico para la generación de tasas medias en intervalos de duración fija [GruCos91]. Los modelos ARIMA(p,d,q) se descomponen en una parte autoregresiva de orden p, una parte integrativa de orden d y una parte de media móvil de orden q. La parte

autoregresiva refleja la dependencia entre la generación actual y las pasadas p generaciones. Así, para un proceso AR(p) los valores generados en una serie temporal $Y=(y_0,y_1,..,y_n)$ se obtienen de los p valores pasados y un factor independiente de la serie temporal, modelable como un proceso de valores idénticamente distribuidos e independientes entre sí $W=(w_0,w_1,..w_n)$. Habitualmente, los valores de la serie W se sintetizan a partir de la realización de una variable aleatoria gaussiana con una media y una desviación típica relacionadas directamente con los correspondientes momentos del proceso AR a generar. De esta forma:

$$y(n)=a_1y(n-1)+a_2y(n-2)+\cdots+a_py(n-p)+w(n) \tag{3.1}$$

donde los términos a_i son coeficientes constantes.

La parte MA(q) del proceso refleja la dependencia en la generación de los valores pasados del proceso independiente que contribuye en el valor obtenido. Así un proceso MA(q) podría expresarse como:

$$x(n)=b_0w(n)+b_1w(n-1)+\cdots+b_qw(n-q) \tag{3.2}$$

donde los términos b_i son coeficientes constantes.

La contribución integrativa, pretende modelar la no estacionariedad de los momentos del proceso estocástico. Si bien podría considerase dentro de la parte AR por su formulación, su síntesis depende de factores distintos de la parte autoregresiva. Así, la parte integrativa también muestra la dependencia con valores pasados de la realización pero depende de los momentos del proceso no estacionario más que de la relación temporal de las generaciones.

El orden d de la parte integrativa queda fijado por el orden del momento del proceso estocástico no estacionario.

En general, se puede expresar la dependencia:

$$z(n)=c_1z(n-1)+c_2z(n-2)+\cdots+c_dz(n-d)+w(n) \tag{3.3}$$

donde:

$$c_i=\binom{d}{i}(-1)^{i+1},\quad i\in\{1,2,\cdots,d\} \tag{3.4}$$

Como caso de aplicación, un proceso cuya media no es estacionaria, pero sí sus momentos de orden superior, tendría una parte integrativa de orden 1. Los procesos integrativos de orden 1 reciben el nombre de marcha aleatoria o "*random walk*" y no están acotados.

La interpretación de un proceso ARIMA(p,d,q) puede ser realizada definiendo el operador de retardo z^{-1} [Pro83]. De esta forma, la expresión general de un proceso ARIMA(p,d,q) quedaría expresada por su transformada *Z* como:

$$S(z) = A^{-1}(z)B(z)C^{-1}(z)W(z) \tag{3.5}$$

Interpretando esta expresión como la relación entre entrada y salida de un filtro digital con excitación *w(n)* y cuya salida es *s(n)* en un instante dado, podríamos definir la función de transferencia del filtro *H(z)* como:

$$H(z) = \frac{B(z)}{A(z)C(z)} \tag{3.6}$$

Obsérvese que las raíces del polinomio *B(z)* se corresponden con los ceros del filtro y los ceros de *A(z)* y *C(z)* con los polos. Según la definición de los valores c_i realizada en la expresión (3.4), el orden integrativo define la multiplicidad del polo en $z=1$, el cual genera la inestabilidad de la respuesta impulsional. El resto de polos z_k, obtenidos a partir de *A(z)*, con $k \in \{1,2,...,p\}$, se encontrarán en el círculo unidad del plano *Z*, es decir, cumplirán $|z_k| < 1$. De forma esquemática se puede representar el modelo ARIMA como se muestra en la figura 3.1.

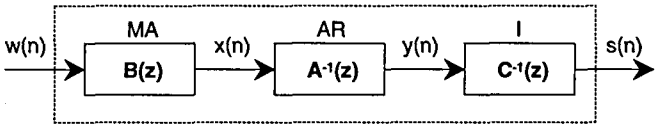

Figura 3.1. Descomposición del modelo ARIMA(p,d,q)

La serie temporal *w(n)*, en general, se denomina serie residual. Se suele considerar como la parte impredecible de la siguiente generación a partir de los valores anteriores. Estos procesos estocásticos son incorrelados y su distribución suele ser gaussiana.

3 Procesos autosemejantes

En esta sección, se hace en primer lugar un repaso de las definiciones y propiedades clásicas de los procesos autosemejantes [Cox84]. Seguidamente, se introducen nuevos resultados que demuestran la existencia de redundancias en las definiciones anteriores [TsyGeo97]. Para ello se considerarán únicamente procesos autosemejantes en tiempo discreto. A continuación, se pondrán algunos ejemplos de procesos que presentan autosemejanza.

3.1 DEFINICIÓN DE LOS PROCESOS AUTOSEMEJANTES

Sea:

$$X = \{X_t\} = (X_1, X_2, \ldots) \tag{3.7}$$

un proceso estocástico estacionario en sentido amplio con parámetro índice (tiempo) discreto. Por tanto, el proceso X tiene un valor medio μ que no depende del tiempo, una varianza σ^2 finita y una función de autocorrelación $r(k)$ que sólo depende de k:

$$\mu = E[X_t]$$

$$\sigma^2 = E\left[(X_t - \mu)^2\right]$$

$$r(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{E\left[(X_t - \mu)^2\right]}, \quad k = 0,1,2,\ldots \tag{3.8}$$

En particular, asumiremos que X presenta una función de autocorrelación de la forma:

$$r(k) \approx k^{-\beta} L_1(k), \quad k \to \infty \tag{3.9}$$

donde $0 < \beta < 1$ y $L_1(k)$ es una función de variación lenta en el infinito, como la función logaritmo, es decir:

$$\lim_{t \to \infty} \frac{L_1(tx)}{L_1(t)} = 1, \quad \forall x > 0 \tag{3.10}$$

Para cada m = 1,2,..., sea $X^{(m)}$ la nueva serie temporal obtenida mediante el promediado de la serie original X en bloques no superpuestos de tamaño m:

$$X^{(m)} = \left(X_k^{(m)} : k = 1,2,\ldots\right)$$

$$X_k^{(m)} = \frac{X_{km-m+1} + \ldots + X_{km}}{m}, \quad k = 1,2,\ldots \tag{3.11}$$

Obsérvese que para cada m, la serie agregada $X^{(m)}$ define un proceso estacionario en sentido amplio. Sea $r^{(m)}(k)$ su función de autocorrelación.

El proceso X es autosemejante exacto de segundo orden con parámetro de Hurst $H=1-\beta/2$ si:

$$\mathrm{var}\left(X^{(m)}\right) = \mathrm{var}(X) m^{-\beta} = \sigma^2 m^{-\beta}, \quad m = 2,3,\ldots \tag{3.12}$$

y:

$$r^{(m)}(k) = r(k), \quad k = 0,1,2,\ldots \tag{3.13}$$

Así, un proceso estacionario X se denomina autosemejante exacto de segundo orden si se cumplen las condiciones (3.9), (3.12) y (3.13).

3.2 DEFINICIÓN REDUCIDA DE LOS PROCESOS AUTOSEMEJANTES

El estudio presentado en [TsyGeo97] permite reducir considerablemente la definición presentada en el apartado anterior para los procesos autosemejantes. En este trabajo se demuestra que las expresiones (3.9), (3.12) y (3.13) no son independientes entre ellas. En primer lugar queda demostrado que un proceso X satisface la condición (3.12) si, y sólo si, su función de autocorrelación es:

$$r(k)=\frac{1}{2}\left[(k+1)^{2-\beta}-2k^{2-\beta}+(k-1)^{2-\beta}\right]\equiv g(k) \qquad 0<\beta<1 \qquad k=1,2,... \tag{3.14}$$

Esta función de autocorrelación se muestra en la figura 3.2 para varios valores de H. Además, la función de variación lenta $L_l(k)$ de la expresión (3.9) deberá ser:

$$\lim_{k\to\infty}\frac{r(k)}{k^{-\beta}}=\frac{1}{2}(2-\beta)(1-\beta)=H(2H-1) \tag{3.15}$$

con H = 1 - β/2, 0 < β < 1.

Finalmente, se demuestra que el proceso X satisface la condición:

$$r^{(m)}(k)=r(k) \qquad k=0,1,... \quad m=2,3,... \tag{3.16}$$

tanto si la expresión (3.12) se cumple como si $r(k) = g(k)$, con lo que se puede dar la siguiente definición para un proceso autosemejante: *El proceso X se denomina autosemejante exacto de segundo orden con parámetro H=1-β/2, 0<β <1, si su función de autocorrelación es la presentada en la expresión (3.14), es decir, si r(k) = g(k).* Cabe remarcar que si en los procesos autosemejantes se cumple que $0 < \beta < 1$ y por tanto $0.5 < H < 1$, en los modelos tradicionales se tenía que $\beta = 1$, lo cual implicaba H=0.5.

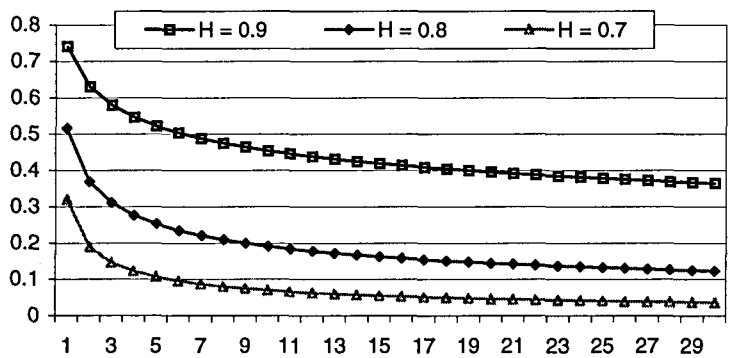

Figura 3.2. Función de autocorrelación para procesos autosemejantes exactos de segundo orden

3.3 PROCESOS ASINTÓTICAMENTE AUTOSEMEJANTES

Un proceso X es asintóticamente autosemejante de segundo orden con parámetro H=1-β/2, 0 < β < 1, si cumple:

$$\lim_{m\to\infty} r^{(m)}(k)=g(k) \tag{3.17}$$

Comparando estos procesos con los exactamente autosemejantes, se observa que para que un proceso sea exactamente autosemejante debe cumplir:

$$r(k) \approx H(2H-1)k^{-\beta} \tag{3.18}$$

mientras que para ser asintóticamente autosemejante es suficiente con que:

$$r(k) \approx ck^{-\beta} \tag{3.19}$$

donde c es una constante que no tiene que ser igual a $H(2H-1)$.

3.4 MODELADO ESTOCÁSTICO DEL FENÓMENO AUTOSEMEJANTE

A menudo se interpretan representaciones que ven los procesos con estructuras de correlación de la forma indicada en la ecuación (3.9) como una suma continua de procesos Gauss-Markov, sugiriendo la presencia de una jerarquía multinivel de mecanismos base para considerar la autosemejanza [LelTaq93]. Sin embargo, es muy difícil en general demostrar la existencia de esta jerarquía al igual que mostrar cómo su presencia contribuye al fenómeno autosemejante. Como resultado se han introducido modelos formales matemáticos que dan paso a elegantes representaciones del fenómeno autosemejante. La mayoría, sin embargo, no dan ninguna interpretación física.

A continuación se presentan dos de estos modelos: el ruido fraccional gaussiano, exactamente autosemejante, y los procesos fraccionales ARIMA, asintóticamente autosemejantes.

3.4.1 Ruido gaussiano fraccional

El movimiento browniano fraccional (*Fractional Brownian Movement*, FBM) es un proceso continuo gaussiano de media nula $B_H=(B_H(s) : s\geq 0)$, $0<H<1$. Este proceso tiene incrementos estacionarios y presenta autosemejanza de parámetro H [LauErr95]. El proceso de incrementos $X_H=(X_H(k)=B_H(k+1)-B_H(k):k\geq 0)$ se denomina ruido gaussiano fraccional (*Fractional Gaussian Noise*, FGN). Éste es un proceso gaussiano estacionario $X=(X_k:k=0,1,2,...)$ con media $\mu=E[X_k]$, varianza $\sigma^2=E[(X_k-\mu)^2]$, y función de autocorrelación:

$$r(k) = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right), \quad k = 1,2,3,... \tag{3.20}$$

Se puede comprobar que, asintóticamente:

$$r(k) \approx H(2H-1)|k|^{2H-2}, \quad 0.5 < H < 1, \quad k \to \infty \tag{3.21}$$

Además, se puede demostrar que los procesos agregados presentan la misma distribución que X para todos los valores de H entre 0 y 1. Por lo tanto, el FGN es un proceso autosemejante exacto de segundo orden de parámetro H, siempre que $1/2<H<1$.

3.4.2 Procesos fraccionales ARIMA (p,d,q)

Un proceso fraccional *ARIMA(p,d,q),* con p y q enteros no negativos y d real, es un proceso estocástico $X=(X_k:k=0,1,2,...)$ con una representación dada por:

$$\Phi(B)\Delta^d X_k = \Theta(B)\varepsilon_k \tag{3.22}$$

donde:

$$\Phi(B)=1-\phi_1 B-\cdots-\phi_p B^p$$
$$\Theta(B)=1-\theta_1 B-\cdots-\theta_q B^q \tag{3.23}$$

son polinomios en el operador de desplazamiento atrás $BX_k=X_{k-1}$, $\Delta=1-B$ representa el operador diferencial, Δ^d es el operador de diferenciación fraccional definido por:

$$\Delta^d=(1-B)^d=\sum_k \binom{d}{k}(-B)^k \tag{3.24}$$

con:

$$\binom{d}{k}(-1)^k=\frac{\Gamma(-d+k)}{\Gamma(-d)\Gamma(k+1)} \tag{3.25}$$

y $(\varepsilon_k:k=0,1,2,...)$ es un proceso de ruido blanco. Según [GraJoy80] y [Cox84] se puede concluir que X es un proceso asintóticamente autosemejante de segundo orden con parámetro $H=d+1/2$, para $0 < d < 1/2$.

Los procesos ARIMA(p,d,q) fraccionales presentan para grandes desplazamientos una función de autocorrelación similar a la de los procesos ARIMA(0,d,0) con el mismo valor de d [Hos81]. Estos últimos son los más simples y fundamentales, presentando una representación de la forma:

$$\Delta^d X_k=\varepsilon_k, \quad k=0,1,2,... \tag{3.26}$$

es decir, una representación autoregresiva de orden infinito. La correspondiente representación de media móvil de orden infinito:

$$X_k=\Delta^{-d}\varepsilon_k \tag{3.27}$$

muestra que los procesos ARIMA(0,d,0) pueden obtenerse sometiendo a un proceso de ruido blanco a una diferenciación fraccional de orden $-d$.

Tomando $H=d+1/2$, tanto el FGN como los procesos ARIMA fraccionales presentan funciones de autocorrelación que se comportan asintóticamente como k^{2d-1}, pero con diferentes constantes de proporcionalidad. Desde el punto de vista de modelado de series temporales, una de las principales ventajas de los procesos ARIMA (0,d,0) sobre el FGN es que los primeros se pueden combinar con los modelos Box-Jenkins [BoxJen76], dando lugar a la familia de procesos ARIMA(p,d,q). Los procesos

ARIMA fraccionales son mucho más flexibles para el modelado de dependencias a corto y largo plazo.

La técnica de diferenciación fraccional presentada en este apartado será utilizada más adelante en este capítulo para la obtención de un nuevo modelo para el tráfico de vídeo MPEG de tasa variable. En ese punto se llevará a cabo también una revisión más detallada de los procesos ARIMA fraccionales y no fraccionales.

4 Generación de procesos con dependencia a largo plazo

Como hemos visto hasta ahora, la LRD es una característica del tráfico que:

- Tiene un impacto práctico en el comportamiento de los sistemas de colas.
- Es de crucial importancia en el dimensionado de dispositivos y en la gestión de la red.
- Si se ignora, las predicciones de comportamiento son demasiado optimistas dando lugar a asignaciones de recursos erróneas.

En los estudios orientados a obtener el comportamiento de redes mediante simulaciones, se necesitarán por tanto generadores de tráfico que incorporen LRD a sus salidas. En esta sección se presentan tres métodos de generación de procesos con LRD. De los dos primeros se pueden encontrar referencias en [HuaDev95], [LauErr95] y [GarWil94]. El tercero es un nuevo método propuesto en este trabajo, que se utilizará posteriormente en este capítulo para la obtención de un modelo para el tráfico de vídeo MPEG de tasa variable a nivel de cuadro [CruPal98][CruAli98b].

4.1 EL ALGORITMO DE HOSKINGS

Este algoritmo es útil para la generación de procesos ARIMA(0,d,0) fraccionales. Las ecuaciones básicas se dan a continuación [Hos84]. El proceso X_k presenta una función de distribución gaussiana con media nula y varianza v_0, y un parámetro de diferenciación fraccional $d=H-1/2$. La función de autocorrelación presenta un comportamiento hiperbólico, y se determina a partir de d del siguiente modo:

$$\rho_k = \frac{d(1+d)\cdots(k-1+d)}{(1-d)(2-d)\cdots(k-d)} \tag{3.28}$$

X_0 se toma de la distribución normal $N(0,v_0)$. Tomando $N_0=0$ y $D_0=1$ se generan n puntos, repitiendo las siguientes acciones para $k = 1, 2, \ldots, n$:

$$N_k = \rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j} \tag{3.29}$$

$$D_k = D_{k-1} - N_{k-1}^2 / D_{k-1} \tag{3.30}$$

$$\phi_{kk} = N_k / D_k \tag{3.31}$$

$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} \qquad j = 1,...,k-1 \tag{3.32}$$

$$m_k = \sum_{j=1}^{k} \phi_{kj} X_{k-j} \tag{3.33}$$

$$v_k = (1-\phi_{kk}^2) v_{k-1} \tag{3.34}$$

Cada X_k se debe escoger de la distribución $N(m_k,v_k)$. Como cada punto depende del anterior, el algoritmo requiere un tiempo de computación de orden $o(n^2)$.

4.2 EL ALGORITMO DE GENERACIÓN RMD

El algoritmo aleatorio del punto medio (*Random Midpoint Algorithm*, RMD) es un algoritmo para la generación aproximada de series FBM [LauErr95].

Supongamos que se quiere generar una serie FBM en el intervalo de tiempo [0,T]. La idea básica del algoritmo RMD es trabajar hacia adentro, subdividiendo el intervalo [0,T] recursivamente y construyendo los valores del proceso en los puntos medios desde los valores en los extremos. La observación clave para el RMD, al construir los valores Z((a+b)/2) en el punto medio del intervalo [a,b], a partir de los valores Z(a) y Z(b) en los extremos, es que si Z fuese verdaderamente FBM, el desplazamiento del punto medio Z((a+b)/2) - (Z(a)-Z(b))/2 sería independiente del incremento Z(b) - Z(a) para todo el intervalo. Este desplazamiento tendría una distribución gaussiana de media nula.

La simplificación clave del RMD, que da lugar a una computación rápida perdiendo exactitud, consiste en escoger todos los desplazamientos independientemente cuando son utilizados en la construcción. Los parámetros que se utilizan en la generación son:

- s_k: desviación estándar utilizada al generar el punto medio en el paso k
- σ_0: desviación estándar del desplazamiento (del FBM) en la escala de tiempo T ($\sigma_0 = T^{2H}$)

Supóngase que $T=2^n$. Por las propiedades de escalado del FBM se tiene que:

$$Var\left[Z\left(\frac{1}{2^k}\right) - Z(0)\right] = \left(\frac{1}{2^k}\right)^{2H} \sigma_0^2 \tag{3.35}$$

Por otra parte:

$$Var\left[Z\left(\frac{1}{2^k}\right) - Z(0)\right] = \frac{1}{4} Var\left[Z\left(\frac{1}{2^{k-1}}\right) - Z(0)\right] + s_k^2 \tag{3.36}$$

Sustituyendo (3.36) en (3.35):

$$\left(\frac{1}{2^k}\right)^{2H} \sigma_0^2 = \frac{1}{4}\left(\frac{1}{2^{k-1}}\right)^{2H} \sigma_0^2 + s_k^2 \tag{3.37}$$

Reordenando:

$$s_k^2 = \left(\frac{1}{2^k}\right)^{2H} \left(1 - 2^{2H-2}\right)\sigma_0^2 \tag{3.38}$$

Y, por lo tanto:

$$s_k = \left(\frac{1}{2^k}\right)^{H} \sqrt{1 - 2^{2H-2}}\,\sigma_0 \tag{3.39}$$

$$s_k = \frac{1}{2^H} s_{k-1} \tag{3.40}$$

El punto inicial es $s_0 = \sqrt{1 - 2^{2H-2}}$. Más específicamente, si el proceso FBM se genera para el intervalo [0,T], se comenzará por fijar Z(0) = 0 y extraer Z(T) de una distribución gaussiana de media nula y varianza T^{2H}. Después, Z(T/2) se calcula como la media de Z(0) y Z(T) mas un offset. Este offset será una variable aleatoria gaussiana con una desviación estándar que vendrá dada por T^{2H} veces el factor de escala inicial:

$$s_1 = 2^{-H} s_0 = 2^{-H}\sqrt{1 - 2^{2H-2}} \tag{3.41}$$

En este punto se reduce el factor de escala por 0.5^H, y los dos intervalos de 0 a T/2 y de T/2 a T se vuelven a dividir. Esta operación se repite hasta el final.

La traza aproximada de FBM (Z(t)) generada por el algoritmo RMD se puede interpretar como el proceso acumulativo de llegadas *A(t)*:

$$A(t) = Mt + \sqrt{aM}\,Z(t) \tag{3.42}$$

donde *M* es la tasa media y *a* es el factor de pico (definido como la relación entre la varianza y la media del número de celdas en una unidad de tiempo entre llegadas). El proceso de incrementos desde el tiempo *t* hasta el *t+1* es por lo tanto:

$$A'(t) = M + \sqrt{aM}\,[Z(t+1) - Z(t)] \tag{3.43}$$

La entrada para la generación de *A'(t)* incluye *M*, *a* y *H*, resultando en un modelo parsimonioso de tráfico con tres parámetros.

Las ventajas de utilizar el algoritmo RMD son su simplicidad, eficiencia y rapidez. La generación de una traza de tráfico FBM con 260000 muestras requiere dos minutos en una estación de trabajo Sun SPARC20. El principal inconveniente consiste en que las trazas no pueden ser generadas en tiempo real, ya que se construye desde los extremos hacia adentro.

4.3 RUIDO BLANCO SOMETIDO A DIFERENCIACIÓN FRACCIONAL

En este apartado se propone y analiza un nuevo método para la generación de procesos con dependencias a largo plazo. Posteriormente, en la sección 6, este método será

utilizado para la obtención de un modelo ARIMA(p,d,q) fraccional para el tráfico de vídeo MPEG VBR.

El método está basado en los procesos de diferenciación fraccional. Estos procesos derivan de los procesos integrativos puros ARIMA(0,d,0), con d real. Un proceso integrativo puro presenta la siguiente función de transferencia [Pro83]:

$$C(z)=\left(1-z^{-1}\right)^{-d} \tag{3.44}$$

Esta función puede desarrollarse en serie de Taylor de la siguiente forma:

$$C(z)=\sum_{i=0}^{\infty} a_i z^{-i} \tag{3.45}$$

donde los coeficientes se obtienen como:

$$a_i=(-1)^i \frac{(-d)(-d-1)\cdots(-d-i+1)}{i!} \tag{3.46}$$

Además, dichos coeficientes pueden relacionarse de forma recursiva de la siguiente manera:

$$\begin{aligned} a_0 &= 1 \\ a_i &= a_{i-1}\frac{(d+i-1)}{i} \end{aligned} \tag{3.47}$$

Por otra parte, estos coeficientes constituyen la respuesta impulsional del filtro ARIMA(0,d,0). En la figura 3.3 se representa esta respuesta impulsional para distintos valores del parámetro de Hurst H. Recuérdese que la relación entre H y d es de la forma:

$$H=d+0.5 \tag{3.48}$$

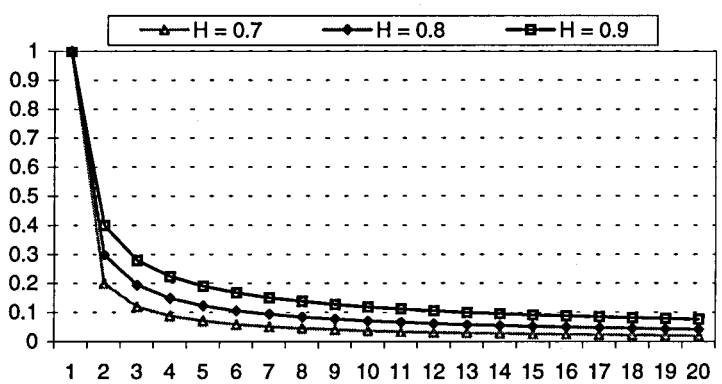

Figura 3.3. Repuesta impulsional del filtro ARIMA(0,d,0)

En la figura 3.4 se muestra la equivalencia entre los filtros propuestos.

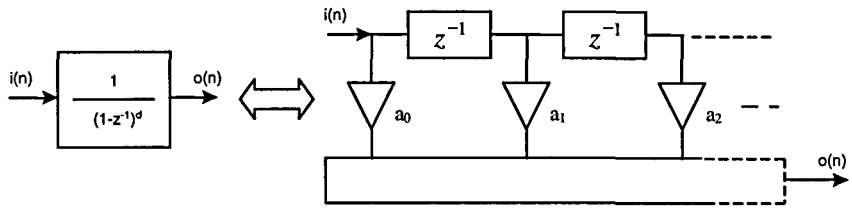

Figura 3.4. Filtro FIR equivalente

El método propuesto en este apartado consiste en pasar un proceso de ruido blanco por el filtro FIR equivalente al ARIMA(0,d,0) tomando un número finito de coeficientes. Uno de los objetivos finales del capítulo será el de dar una cota mínima para el número de coeficientes que deben utilizarse para garantizar el grado de LRD deseada en las generaciones de tráfico.

Para obtener una validación inicial de este método de generación se puede proceder de la siguiente forma. Sea *i(n)* un proceso de ruido blanco a la entrada del filtro FIR equivalente. La salida *o(n)* se obtendrá mediante la convolución de la serie de entrada con la respuesta impulsional del filtro *h(n)*:

$$o(n) = i(n) * h(n) \tag{3.49}$$

Por tanto, las funciones de autocorrelación de estos procesos deben satisfacer:

$$r_{oo}(n) = r_{ii}(n) * r_{hh}(n) \tag{3.50}$$

donde $r_{xx}(n)$ representa la función de autocorrelación del proceso $x(n)$. El proceso de entrada está incorrelado, y por tanto:

$$r_{ii}(n) = \delta(n) \tag{3.51}$$

donde $\delta(n)$ es la función delta de Dirichlet. Entonces:

$$r_{oo}(n) = r_{hh}(n) \tag{3.52}$$

Así, la función de autocorrelación de la salida del filtro ha de ser la misma que la de los coeficientes dados en la expresión (3.46). Además, la salida del filtro ha de ser un proceso autosemejante de segundo orden, con lo que su función de autocorrelación debe satisfacer la expresión (3.14). Con el objetivo de validar este estudio, en la figura 3.5 se compara la autocorrelación de 50000 coeficientes según la expresión (3.46) con la autocorrelación de la expresión (3.14), para H=0.8 (d=0.3, β=0.4), mostrándose la bondad del ajuste.

Como última prueba, se pasó una secuencia de 50000 muestras de ruido gaussiano blanco por el filtro propuesto, con diferentes valores del parámetro *d*. El número de coeficientes tomados para el filtro fue igualmente de 50000. Las funciones de autocorrelación de las salidas, comparadas con la de la entrada, se presentan en la figura 3.6. Los valores tomados para el parámetro *d* fueron 0.2, 0.3 y 0.4, correspondiendo a valores de *H* de 0.7, 0.8 y 0.9 respectivamente. En la parte izquierda de la figura se

muestra un número pequeño de desplazamientos, mientras que en la parte derecha este número se hace mayor para comprobar la presencia de LRD.

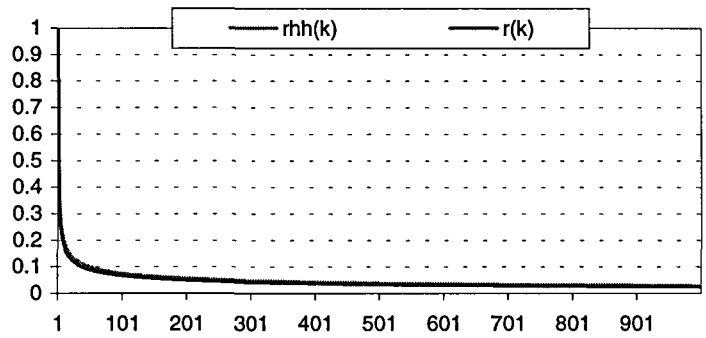

Figura 3.5. Comparación de las funciones de autocorrelación

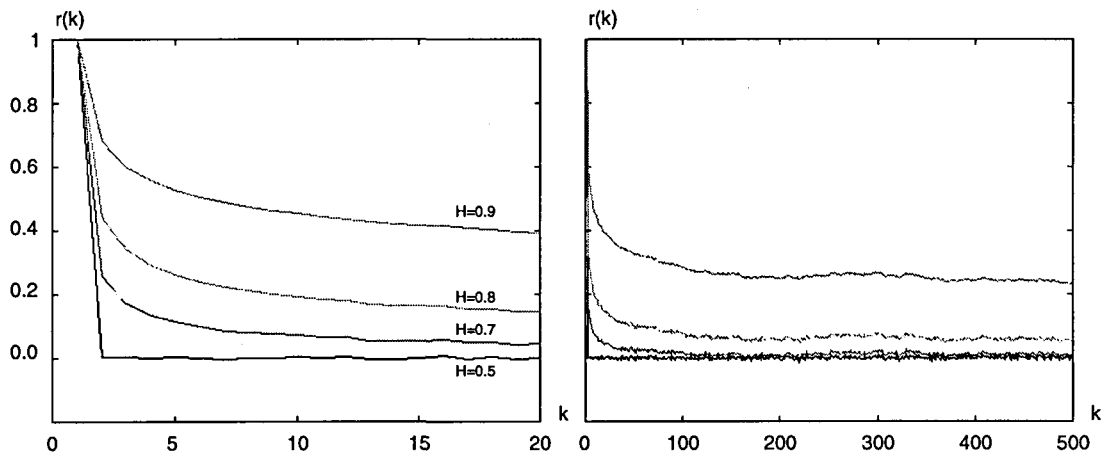

Figura 3.6. Funciones de autocorrelación a la entrada y a la salida del filtro FIR

El método presentado para la generación de procesos con LRD presenta tres claras ventajas con respecto a los anteriores:

- En primer lugar, el coste de computación es despreciable y constante para cada nueva generación, en contraposición con el algoritmo de Hoskings que presentaba un orden de computación elevado y creciente geométricamente para cada nueva generación.
- No se hace necesario el cálculo de la serie completa antes de su utilización, como en el caso del RMD.
- Además, el método trabaja sin problemas en tiempo real, con lo que sus generaciones pueden ser utilizadas para conducir simulaciones.

En la sección 6 de este capítulo se complementará este modelo con dos nuevos filtros, una autoregresivo y otro de media móvil, para obtener un modelo adecuado para el tráfico de vídeo MPEG de tasa variable.

5 Análisis del grado de autosemejanza de los procesos

Un proceso estocástico que satisfaga la ecuación (3.9) se dice que exhibe LRD. Por tanto, los procesos con LRD están básicamente caracterizados por una función de autocorrelación que decae hiperbólicamente a medida que el retardo aumenta. Además, es sencillo ver que la expresión de la ecuación (3.9) implica que:

$$\sum_k r(k) \to \infty \tag{3.53}$$

Esta divergencia de la suma de los valores de la autocorrelación captura la idea que hay detrás de la LRD: mientras correlaciones con grandes retardos son individualmente pequeñas, su efecto acumulativo es importante y comporta características que son drásticamente diferentes de las de los procesos más convencionales, los cuales presentan tan sólo dependencias a corto plazo. Los procesos que únicamente presentan SRD están caracterizados por un decaimiento exponencial de las funciones de autocorrelación:

$$r(k) \approx \rho^k, \quad k \to \infty, \quad 0 < \rho < 1 \tag{3.54}$$

resultando en una función de autocorrelación sumable:

$$0 < \sum_k r(k) < \infty \tag{3.55}$$

Cuando se trabaja en el dominio de la frecuencia, la LRD se manifiesta en el comportamiento potencial de la función de densidad espectral de potencia en el origen:

$$f(\lambda) = \sum_{k=0}^{\infty} r(k) e^{j\lambda k} \approx \lambda^{-\gamma} L_2(\lambda), \quad \lambda \to 0 \tag{3.56}$$

donde $0 < \gamma < 1$ y $L_2(\lambda)$ es una función que varía lentamente en las cercanías del origen. Así, desde el punto de vista frecuencial, la LRD implica la divergencia de la función de densidad espectral de frecuencia en el origen. Sin embargo, la SRD está caracterizada por una función de densidad espectral de potencia positiva y finita para $\lambda = 0$.

Hasta este punto se han estudiado, por una parte, las distintas definiciones de los procesos autosemejantes, y por otra algunos métodos de generación de procesos con LRD. En este apartado se hablará de las técnicas existentes para el análisis del grado de autosemejanza que presenta un proceso, en particular mediante la estimación del parámetro *H*. Dichas técnicas se basan en la búsqueda de LRD o de funciones de densidad de probabilidad abiertas, y en presencia de alguna de las dos buscar sus parámetros. Para aplicarlas, se hace necesario conocer cuáles son sus puntos débiles y fuertes, así como sus requisitos y sus posibles errores. En concreto, se debe evitar la aplicación a ciegas de alguno de los métodos conocidos que podría llevar a interpretaciones erróneas. En general, es difícil estimar con exactitud el parámetro *H*

definitorio de la autosemejanza de un proceso, siendo conveniente confirmar los resultados mediante la utilización de más de un método.

Existen dos grupos de técnicas, unas más heurísticas y otras más rigurosas. Entre las primeras se encuentran:

- Análisis del decaimiento de las varianzas
- Análisis de la estadística R/S
- Periodograma
- Correlograma

El segundo grupo de técnicas es más riguroso, proporcionando intervalos de confianza. Entre ellas se encuentran las siguientes:

- Estimación de máxima verosimilitud de Whittle
- Análisis semiparamétrico de Robinson
- Técnicas *wavelet*

El primer grupo de técnicas ha sido hasta la actualidad el más comúnmente utilizado. A continuación se exponen cada una de ellas. Además es de destacar que, dentro del segundo grupo de técnicas, las basadas en *wavelets* son cada vez más utilizadas y aceptadas.

5.1 MÉTODOS ESTADÍSTICOS PARA EL ANÁLISIS DE LA LRD

5.1.1 LRD y el efecto de Hurst

Históricamente, la importancia de los procesos autosemejantes reside en el hecho de que proporcionan una explicación e interpretación elegante de una ley empírica que se conoce con el nombre de *ley de Hurst* o *efecto Hurst*. Dicho efecto se introduce brevemente a continuación. Para un conjunto de observaciones dado (X_k: k = 1, 2,..., n), con media $X_m(n)$ y varianza $S^2(n)$, la estadística *R/S* viene dada por:

$$\frac{R(n)}{S(n)} = \frac{max(0, W_1, W_2, ..., W_n) - min(0, W_1, W_2, ..., W_n)}{S(n)} \tag{3.57}$$

con:

$$W_k = (X_1 + X_2 + ... + X_k) - kX_m(n), \quad k = 1,2,...,n \tag{3.58}$$

Hurst encontró que un gran número de series temporales naturales estaban bien representadas por la relación:

$$E\left[\frac{R(n)}{S(n)}\right] \approx cn^H, \quad n \to \infty \tag{3.59}$$

con parámetro de Hurst H típicamente alrededor de 0.73, y c una constante positiva que no depende de n. Por otra parte, si las observaciones X_k provienen de un proceso SRD, se cumple

$$E\left[\frac{R(n)}{S(n)}\right] \approx dn^{0.5}, \quad n \to \infty \tag{3.60}$$

donde d es una constante finita e independiente de n. Esta discrepancia se conoce generalmente con el nombre de efecto Hurst.

El objetivo del análisis R/S es encontrar el grado de autosemejanza H del proceso que ha generado el conjunto de muestras en consideración. En la práctica, este análisis se basa en un enfoque gráfico heurístico que intenta explotar al máximo la información que pueda dar el conjunto de muestras.

Dado un conjunto de N observaciones (X_k: k = 1, 2, ..., N), se subdivide en K bloques que no se sobrepongan, y se calcula $R(t_i,n)/S(t_i,n)$ para cada uno de los nuevos puntos iniciales:

$$t_1 = 1; \quad t_2 = \frac{N}{K} + 1; \quad t_3 = \frac{2N}{K} + 1; \quad \ldots \tag{3.61}$$

que satisfagan $(t_i\text{-}1) + n \leq N$.

Aquí, la estadística $R(t_i,n)/S(t_i,n)$ está definida como en (3.57), sustituyendo W_k por $(W_{ti+k}\text{-}W_{ti})$, y siendo $S^2(t_i,n)$ la varianza de $\{X_{ti+1}, X_{ti+2}, \ldots, X_{ti+n}\}$. Así, para un valor dado de n, se obtienen varias muestras de R/S. Si n es pequeño, el número de muestras será próximo a K, mientras que para n grande, dicho número tenderá a la unidad. Posteriormente, se van tomando valores de n logarítmicamente espaciados y se representa gráficamente el logaritmo de $R(t_i,n)/S(t_i,n)$ en función del logaritmo de n, obteniendo un gráfico de puntos dispersos. De este modo se puede encontrar una estimación del parámetro de Hurst H', aproximando el gráfico por una recta mediante la técnica de errores cuadráticos mínimos. La pendiente de esta recta es directamente la estimación buscada H'.

5.1.2 Análisis del decaimiento de las varianzas

Desde un punto de vista estadístico, la propiedad más notable de los procesos autosemejantes es que la varianza de la media aritmética se comporta como $n^{-\beta}$ para alguna $\beta \in (0,1)$, en lugar de n^{-1} como se tiene con los procesos para los cuales las series agregadas tienden a ruido puro de segundo orden.

En [TsyGeo97] queda demostrado que un proceso es autosemejante exacto de segundo orden si se cumple:

$$\text{var}\, X^{(m)} = \sigma^2 m^{-\beta} \tag{3.62}$$

donde σ^2 representa la varianza del proceso original X.

Esta propiedad da lugar al método de análisis de las varianzas como método de obtención del grado de LRD de un proceso en particular. Para ello se lleva a cabo la representación gráfica de las varianzas, representando en un eje el logaritmo de $var(X^{(m)})$ y en el otro el logaritmo de m. Al igual que en la técnica R/S, el gráfico debe ajustarse por una recta mediante una aproximación de errores cuadráticos mínimos. Este ajuste deberá realizarse para valores grandes de m, que es cuando la representación tiende a una recta. Valores estimados de la pendiente β' entre -1 y 0 sugieren autosemejanza. El grado de autosemejanza obtenido viene dado por $H'=1-\beta'/2$.

5.1.3 Análisis del periodograma

Este análisis se basa en el comportamiento de la función de densidad espectral de potencia del proceso bajo estudio en las cercanías del origen. Como se ha comentado anteriormente, en ausencia de LRD se encontrará una función plana, mientras que los procesos con LRD presentan una función de densidad de potencia divergente en el origen.

Para encontrar una estimación del parámetro de Hurst, se calcula en primer lugar el periodograma, se toman los valores de bajas frecuencias y, tras una conversión logarítmica, se aproximan los valores por una recta utilizando errores cuadráticos mínimos. Con el valor de la pendiente α se puede obtener una estimación del parámetro de Hurst mediante $H' = (-\alpha+1)/2$.

5.1.4 Análisis del correlograma

Este método consiste simplemente en la representación de la función de autocorrelación en función del desplazamiento. Para grandes desplazamientos, se obtendrá una pendiente de dicha función de valor $-\beta$, estimándose un parámetro de Hurst $H'=1-\beta/2$.

5.2 ANÁLISIS DE TRAZAS DE TRÁFICO REAL Y ARTIFICIAL

En este apartado se analizarán, mediante tres de las técnicas expuestas anteriormente, una serie de trazas tanto de tráfico real como de tráfico sintético generado mediante el proceso de diferenciación fraccional introducido en el apartado 4 de este capítulo.

En primer lugar se lleva a cabo el análisis de diferentes secuencias generadas artificialmente. Para ello se ha ido variando la longitud de la secuencia generada y el número de coeficientes utilizados en el filtro FIR equivalente. Los resultados se muestran en la tabla 3.2. Como primera conclusión se puede afirmar que los tres métodos utilizados son bastante buenos para el análisis de este tipo de secuencias. De entre ellos, el método de las varianzas es el más sensible a los cambios en el número de

coeficientes utilizados. Fijándonos en los resultados obtenidos con este método, al ser el más restrictivo, podemos concluir que un número de coeficientes entre 5000 y 10000 para el filtro utilizado es suficiente para asegurar el grado de LRD buscado. Por otra parte, es de destacar como los resultados se alejan más de los esperados al crecer el parámetro *H*.

Longitud secuencia generada (cuadros)	Número de coeficientes	H nominal	Periodograma	Varianzas	R/S
50000	50000	0.7	0.696	0.686	0.692
		0.8	0.792	0.786	0.773
		0.9	0.896	0.876	0.845
50000	10000	0.7	0.697	0.685	0.693
		0.8	0.797	0.782	0.774
		0.9	0.897	0.867	0.848
50000	5000	0.7	0.698	0.681	0.692
		0.8	0.798	0.772	0.775
		0.9	0.897	0.852	0.851
50000	1000	0.7	0.698	0.671	0.689
		0.8	0.798	0.751	0.768
		0.9	0.897	0.812	0.837
100000	50000	0.7	0.695	0.709	0.722
		0.8	0.795	0.805	0.806
		0.9	0.894	0.89	0.88
100000	10000	0.7	0.695	0.709	0.722
		0.8	0.795	0.804	0.807
		0.9	0.894	0.88	0.882
100000	5000	0.7	0.694	0.704	0.722
		0.8	0.794	0.794	0.808
		0.9	0.894	0.868	0.882
100000	1000	0.7	0.694	0.677	0.709
		0.8	0.794	0.748	0.784
		0.9	0.893	0.806	0.848

Tabla 3.2. Análisis de tráfico artificial

Por otra parte, en la tabla 3.3 se presentan los resultados obtenidos en el análisis de trazas de tráfico de vídeo real, utilizando en este caso sólo el método de varianzas. Los parámetros utilizados son los mismos que los utilizados para el análisis de trazas de tráfico artificial, y se validan en primer lugar mediante el análisis de secuencias incorreladas. De los resultados obtenidos se comprueba la presencia de LRD en el tráfico de vídeo MPEG VBR, con valores típicos de *H* sobre 0.8 y 0.9.

Así, de los dos grupos de medidas llevados a cabo, es posible concluir por una parte el correcto funcionamiento del método propuesto para la generación de tráfico con LRD. Además, el método de las varianzas es el más sensible al cambio en el número de coeficientes utilizado, con lo cual se convierte en el método idóneo para el análisis del parámetro *H* en el contexto de este trabajo.

Secuencia	Sistema	Cuadros	Varianza
Ruido Gaussiano		50000	0.491
Ruido Gaussiano		150000	0.482
Jurassic Park	PAL	175752	0.868
América	PAL	34799	0.912
Geografía Cat.	PAL	51000	0.903
Neil Young	NTSC	47981	0.909
Blade Runner	NTSC	156431	0.89
Midnight Murders	NTSC	39935	0.831

Tabla 3.3. Análisis de trazas de ruido y tráfico real

6 Modelo ARIMA fraccional para tráfico de vídeo VBR MPEG a nivel de cuadro

Los modelos ARIMA(p,d,q) han sido ampliamente utilizados en varios campos de investigación. Estos modelos ajustan las funciones de autocorrelación del tráfico generado por la codificación MPEG VBR. Presentan como inconveniente que, sin ningún acondicionamiento, pueden generar eventos negativos. Además, las funciones de distribución de probabilidad generadas necesitan en ocasiones ser reajustadas mediante alguna función de transformación, con el fin de aproximarse a la función de distribución de la serie a modelar.

En trabajos anteriores [Mat96][CruAli97] se obtuvo un modelo ARIMA no fraccional para el tráfico de vídeo a nivel de cuadro. El principal inconveniente de este modelo es que presenta una parte integrativa no nula y no fraccional, con lo que no puede ser utilizado para la síntesis artificial de tráfico, al no estar acotada su salida. Sin embargo, sus características como predictor a corto plazo son excelentes, y será utilizado precisamente con este fin en el capítulo 4 de este trabajo.

En este apartado, sin embargo, se obtendrá un modelo ARIMA(p,d,q) cuya parte integrativa será fraccional, con lo que su tasa de salida estará acotada. Así, el modelo será útil para la generación de tráfico de vídeo a nivel de cuadro. El modelo se basará, en su parte integrativa, en el método de diferenciación fraccional desarrollado en el apartado 4.

Como se estudió en dicho apartado, la parte integrativa presenta una función de transferencia que se puede desarrollar dando lugar a un filtro FIR de infinitos coeficientes. En este apartado se tomará un número finito de coeficientes para la obtención del modelo, y finalmente se obtendrá un margen de valores para dicho número de coeficientes que nos proporcionará la LRD deseada.

Para el ajuste del modelo se han utilizado las tres secuencias que se presentan en la tabla 3.4, obteniéndose resultados extremadamente parecidos para todas ellas. Además,

se detallan los parámetros N (tamaño GoP), M (tamaño SGoP) y Q (paso de cuantificación) correspondientes al codificador MPEG.

Nombre	Número de cuadros	N	M	Q	Sistema
Jurassic Park	175752	6	2	6	PAL
América	34799	6	2	6	PAL
Midnight murders	39935	6	2	6	NTSC

Tabla 3.4. Secuencias utilizadas para el ajuste del modelo

El primer paso a dar en la obtención del modelo es la extracción de la parte integrativa de las serie temporal originales. Para ello se utilizará precisamente el filtro inverso al considerado en el apartado 4 para la obtención de las dependencias a largo plazo. Es decir, si tenemos, fijándonos en la figura 3.1:

$$S(z) = Y(z)C^{-1}(z) \tag{3.63}$$

podremos obtener:

$$Y(z) = S(z)C(z) \tag{3.64}$$

Obsérvese que en éste momento se está tomando la definición de la función de transferencia tal y como aparece en la figura 3.1, es decir:

$$C^{-1}(z) = (1-z^{-1})^{-d} \Rightarrow C(z) = (1-z^{-1})^{d} \tag{3.65}$$

Esta expresión se puede de nuevo desarrollar en serie de Taylor, dando lugar a un filtro FIR de infinitos coeficientes. La tendencia de dichos coeficientes se puede ver en la figura 3.7, para un valor del parámetro d igual a 0.3 ($H = 0.8$).

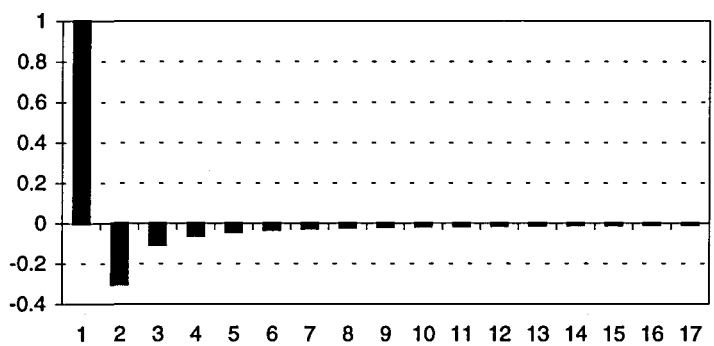

Figura 3.7. Coeficientes del filtro inverso C(z)

Para la extracción de la parte integrativa, se utilizó un número de coeficientes variable entre 10000 y 50000, siendo los resultados prácticamente idénticos en todos los casos.

Posteriormente, una vez que se dispone de la serie original sin su componente integrativa, $y(n)$, se procede al ajuste de la parte autoregresiva del modelo. Esta componente se encarga fundamentalmente de modelar la dependencia de la generación actual del modelo con las generaciones pasadas.

La serie temporal *y(n)* presenta un comportamiento estacional de periodo *N*, precisamente el valor escogido para el tamaño del grupo de cuadros en la codificación MPEG. Este comportamiento se refleja en su función de autocorrelación, como puede observarse en la figura 3.8. Utilizando los picos de esta función de autocorrelación, que aparecen en los múltiplos de *N*, la componente autoregresiva puede ser sintetizada. Para determinar sus coeficientes se empleó el método de mínimos cuadrados, obteniéndose un buen ajuste con un filtro autoregresivo de orden 2. Realizando la extensión necesaria del filtro debido a que sólo se habían tenido en cuenta los picos de la función de autocorrelación, la expresión para la componente AR del modelo buscado queda de la siguiente manera:

$$A(z) = (1 - 0.626z^{-6} - 0.342z^{-12}) \tag{3.66}$$

Finalmente, se hace necesario el ajuste de la componente de media móvil del modelo. Para ello, en primer lugar se extrae la componente autoregresiva de la serie y(n) según:

$$X(z) = A(z)Y(z) \tag{3.67}$$

obteniéndose la serie *x(n)*. Para estimar los parámetros de la componente MA, se utilizó de nuevo el método de mínimos cuadrados para el ajuste de la función de autocovarianza de *x(n)*. El mejor ajuste se consiguió con un filtro de orden 13 con la siguiente función de transferencia asociada:

$$\begin{aligned} B(z) = {} & 1 - 0.0751z^{-1} + 0.221z^{-2} + 0.0039z^{-3} + 0.1816z^{-4} - 0.0149z^{-5} \\ & - 0.0599z^{-6} - 0.0254z^{-7} + 0.0500z^{-8} - 0.0286z^{-9} + 0.0189z^{-10} \\ & - 0.0260z^{-11} - 0.1762z^{-12} - 0.0116z^{-13} \end{aligned} \tag{3.68}$$

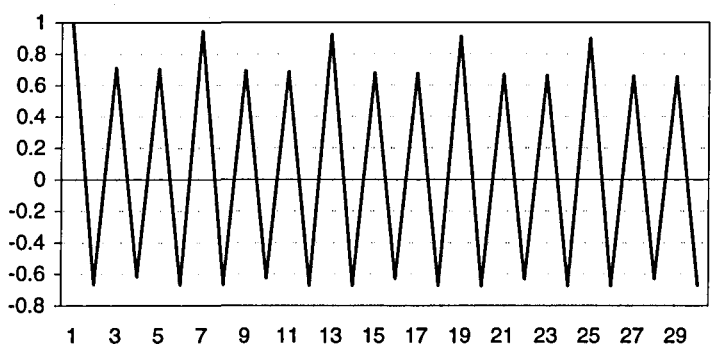

Figura 3.8. Función de autocorrelación de la serie y(n)

Finalmente, la función de transferencia del filtro completo se obtiene como:

$$H(z) = \frac{B(z)}{A(z)C(z)} \tag{3.69}$$

Con objeto de evaluar el comportamiento de dicha función de transferencia, un análisis de los errores de predicción se realizó para todas las secuencias. En la figura 3.9 se representa la autocorrelación de los residuos (errores de predicción) y el intervalo de confianza al 99 % para la secuencia "América", observándose como los errores de predicción están realmente incorrelados. Por lo tanto, el modelo obtenido permite una buena predicción de la tasa de salida de un codificador VBR MPEG a nivel de cuadro.

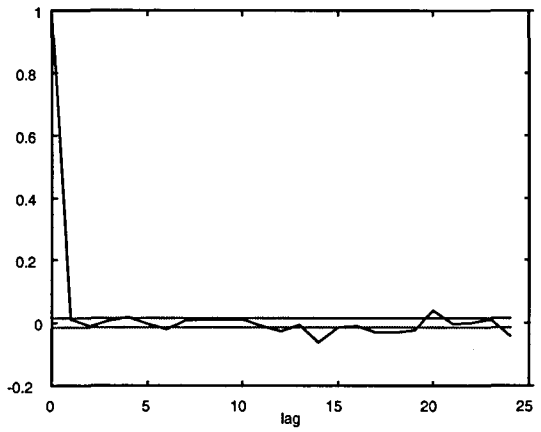

Figura 3.9. Función de autocorrelación de los residuos

El modelo obtenido es utilizado ahora para la generación de tráfico sintético, que pueda ser utilizado para conducir simulaciones. El objetivo principal es de nuevo la obtención de una cota mínima para el número de coeficientes necesitados en el filtro integrativo para obtener el grado de LRD deseado.

De esta forma, a la entrada del filtro se aplicó un proceso de ruido gaussiano blanco, con media nula y desviación estándar igual a la unidad. El número de coeficientes, de la parte integrativa se fue variando en el margen comprendido entre los 100 y los 50000 coeficientes. Al igual que en la etapa de construcción del modelo, el valor de d tomado para la parte integrativa fue de 0.3, que corresponde a un valor de 0.8 para el parámetro de Hurst. Mediante el test de las varianzas se analizó el grado de LRD obtenido. Los resultados se presentan en la tabla 3.5. Con 10000 coeficientes, el parámetro H medido es de 0.782, suficientemente cercano ya al valor de 0.8 esperado. Incluso 5000 coeficientes podrían ser suficientes para algunas aplicaciones.

Con objeto de utilizar las trazas obtenidas para conducir simulaciones, el valor medio y la varianza de las series generadas debe ajustarse a los valores típicos presentados por el tráfico de vídeo VBR MPEG. En la figura 3.10 se muestra el esquema de bloques final. Gracias al factor de amplificación D se pueden modelar secuencias con distintos grados de variabilidad. Por otra parte, variando la media de la serie generada sumando un valor m, se ajustan distintos niveles de calidad del codificador MPEG.

Nº coef.	H
100	0.598
500	0.657
1000	0.714
2000	0.747
3000	0.761
4000	0.767
5000	0.771
10000	0.782
25000	0.783

Tabla 3.5. Relación entre el número de coeficientes del filtro integrativo y el parámetro de Hurst

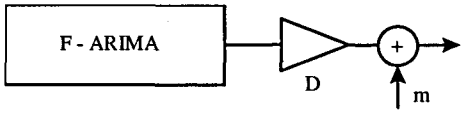

Figura 3.10. Diagrama de bloques del generador de tráfico.

Para llevar a cabo un primer ajuste práctico, se trabajó con una secuencia real de aproximadamente 225000 cuadros, resultante de la concatenación de las secuencias "Jurassic Park" y "Geografía de Cataluña". Esta secuencia presentó un valor medio de 47533 bits por cuadro y una desviación estándar de 36219 bits. Una vez ajustadas con estos valores, se obtuvo un nuevo conjunto de trazas de tráfico artificial. En la figura 3.11 se representa la evolución temporal de una de ellas, en particular la obtenida utilizando 10000 coeficientes para el filtro integrativo.

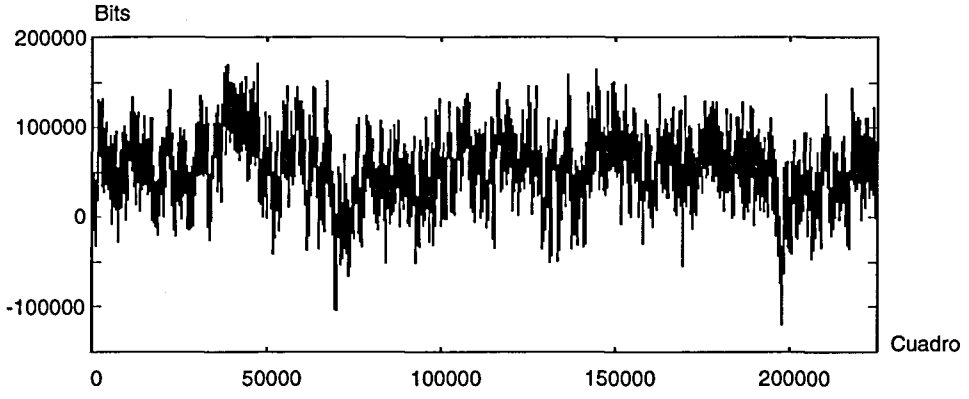

Figura 3.11. Representación temporal de la traza generada con 10000 coeficientes

Como se observa en la figura, la traza generada proporciona en ocasiones valores negativos, que evidentemente no pueden ser utilizados como valores válidos para representar el número de bits por cuadro de una secuencia de vídeo. La distribución de esta secuencia de valores es prácticamente una distribución normal. Una posible solución consiste en ajustar el valor mínimo de la serie al valor mínimo de la secuencia original. En el caso de la serie concatenada, este valor es de 4383 bits por cuadro. En la figura 3.12 se representa la serie rectificada. Un ajuste más profundo pasaría por la proyección de la función de distribución obtenida sobre una función de distribución que capture más correctamente la presentada por el tráfico real. Entre estas funciones se

encuentran la función lognormal, o bien una combinación de las funciones gamma y Pareto [GarWil94].

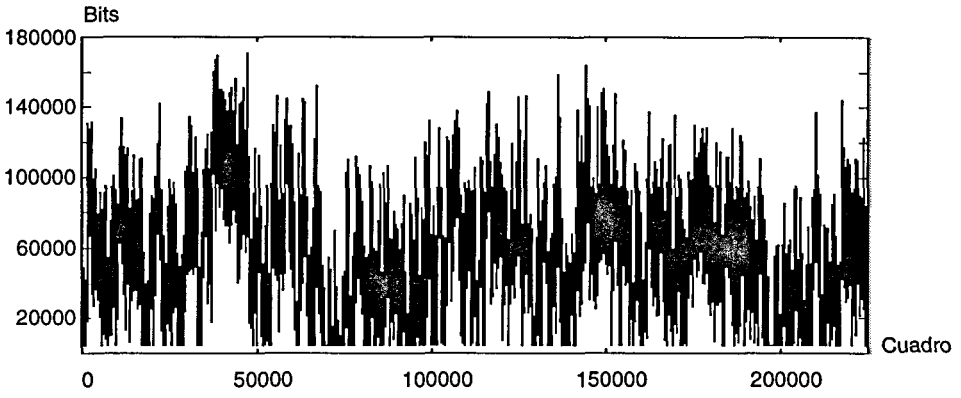

Figura 3.12. Representación temporal de la traza generada y rectificada

Las funciones de autocorrelación de la serie original y de las dos trazas sintéticas se muestran y comparan en la figura 3.13. Por motivos de claridad, el comportamiento estacional presentado por las tres funciones ha sido previamente extraído. La pendiente para desplazamientos grandes, la cual está relacionada con el grado de LRD, es muy parecida en los tres casos, lo cual permite comprobar la bondad del resultado obtenido.

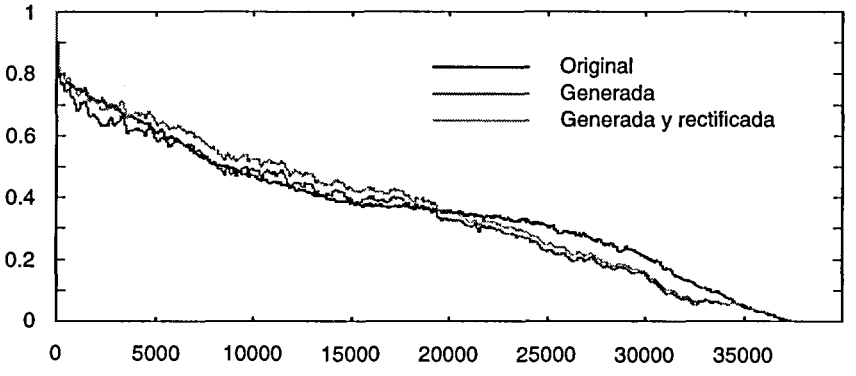

Figura 3.13. Comparación de las funciones de autocorrelación de las secuencias original y sintetizada

7 Conclusiones

En este capítulo se ha realizado un estudio de los procesos autosemejantes, con aplicación directa sobre el modelado del tráfico de vídeo MPEG VBR. Para ello, en primer lugar se ha introducido un repaso sobre la definición, características y propiedades de los procesos autosemejantes. Entre las características de estos procesos, destaca la presencia de dependencias a largo plazo, que no son capturadas correctamente por los modelos de tráfico convencionales. El grado de autosemejanza de un proceso se caracteriza mediante el parámetro de Hurst. El dimensionado de capacidades de enlaces

y de longitudes de buffers de almacenamiento en nodos ATM es, en general, demasiado optimista cuando no se tiene en cuenta la LRD.

A continuación se han presentado dos de los métodos clásicos para generar artificialmente tráfico con un grado de autosemejanza determinado. Estos métodos presentan unos inconvenientes claros en función de la cantidad de memoria y de tiempo que necesitan para ejecutarse. Esto los hace especialmente inadecuados cuando se trata de utilizarlos para generar la entrada de simulaciones en tiempo real. En este trabajo se ha propuesto un nuevo método, basado en los procesos de diferenciación fraccional, que si permite la generación de tráfico autosemejante en tiempo real, sin necesidad de grandes cantidades de memoria. La implementación exacta del método implicaría la necesidad de un filtro de infinitos coeficientes para imitar el comportamiento de un filtro integrativo puro. Sin embargo, tras los experimentos realizados, se comprobó como era suficiente con la utilización de un filtro FIR, y por tanto con un número finito de coeficientes. Uno de los objetivos de los experimentos llevados a cabo posteriormente fue el de obtener el número mínimo de coeficientes necesarios para dicho filtro que proporcionarían el grado buscado de autosemejanza.

La validez del método presentado se ha verificado de varias formas. Por una parte, se ha comprobado analítica y gráficamente como la función de autocorrelación del tráfico generado presenta las características esperadas según los resultados de trabajos previos. Además, se han introducido las técnicas clásicas de análisis de procesos autosemejantes. Mediante su utilización, se ha verificado de nuevo como el tráfico generado presentaba el grado de LRD deseado. Además, se pudo obtener la cota buscada para el mínimo número de coeficientes necesario en el filtro FIR, resultando estar entre los 5000 y 10000 coeficientes en función de la exactitud deseada para el parámetro H obtenido.

A continuación, se ha empleado dicho método para la obtención de un modelo ARIMA fraccional para el tráfico de vídeo MPEG VBR a nivel de cuadro. El modelo se realizó a partir de varias trazas suficientemente largas de tráfico real, obteniéndose un correcto ajuste de los residuos de predicción para cada una de ellas. Como resultado, se obtuvieron unos coeficientes muy parecidos para los filtros de media móvil y autoregresivo de cada una de las secuencias, permitiendo proponer un modelo común para todas ellas. Finalmente, el modelo ha sido utilizado para la generación artificial de tráfico de vídeo. De nuevo, se varió el número de coeficientes en el filtro integrativo, obteniéndose los mismos resultados para el valor mínimo necesitado que ofrecía el grado de autosemejanza clásico de las secuencias de vídeo. Además, se comprobó el parecido de la función de autocorrelación con la presentada por el tráfico real. Por otra parte, la función de densidad de probabilidad presentada por la secuencia de salida del

modelo es una función normal. Como último ajuste, esta función debería ser proyectada a alguna de las funciones de densidad que se acercan más a las del tráfico real, como la función lognormal o una combinación de las funciones gamma y Pareto.

CAPÍTULO 4

CONFORMACIÓN DE FUENTE PARA EL TRÁFICO DE VÍDEO MPEG VBR

Los algoritmos utilizados en la actualidad para la compresión de tráfico de vídeo, como por ejemplo el MPEG, provocan fluctuaciones periódicas de la tasa generada a la salida del codificador. Estas fluctuaciones se producen como resultado de los distintos modos de codificación empleados por el algoritmo en los distintos cuadros de la secuencia de vídeo. Al transmitir sobre redes ATM, la ubicación de recursos se puede ver afectada negativamente, debido a que la periodicidad comentada provoca un descenso en la ganancia de multiplexación estadística. Para evitar estos efectos, se recurre al uso de técnicas de suavizado.

En el presente capítulo se exponen y analizan las técnicas clásicas de suavizado, así como algunas más recientes, estudiando sus ventajas e inconvenientes. Entre estos últimos se encuentra el retardo adicional introducido, inadmisible en servicios interactivos. Como solución se presenta un nuevo conformador, basado en técnicas de predicción, que consigue la máxima decorrelación de la serie de salida minimizando el retardo por debajo de una cota previamente establecida. El predictor utilizado por este conformador se construye en base al modelo ARIMA no fraccional para la tasa a nivel de cuadro previamente desarrollado. Finalmente, se comprueba la ganancia introducida al transmitir el tráfico sobre redes ATM y se analizan casos particulares de multiplexación.

1 Introducción

En el capítulo anterior se ha caracterizado y modelado el tráfico de vídeo MPEG VBR a nivel de cuadro. Con el estudio realizado, se ha alcanzado un grado de conocimiento suficiente de este tipo de tráfico, lo cual nos permite plantearnos un nuevo problema: cuál debe ser la forma óptima de entregar este tráfico a una red ATM. Como ya se ha comentado anteriormente, una de los objetivos fundamentales a la hora de trabajar con la B-ISDN es la utilización eficiente de los recursos de red. Así, la bondad de la compartición de canales de transmisión, cuantificada en términos de ganancia de multiplexación estadística, ha de ser lo más elevada posible.

Sin embargo, a la hora de transmitir tráfico de vídeo MPEG VBR aparece un serio problema. Los algoritmos utilizados en este estándar provocan fluctuaciones periódicas de la tasa generada a la salida del codificador. Estas fluctuaciones se producen como resultado de los distintos modos de codificación empleados por el algoritmo en los distintos cuadros de la secuencia de vídeo. Al transmitir sobre redes ATM, la ubicación de recursos se puede ver afectada negativamente, debido a que la periodicidad comentada provoca un descenso en la ganancia de multiplexación estadística [Kar96]. Para evitar estos efectos, se recurre al uso de técnicas de suavizado.

La más simple de dichas técnicas consiste en el almacenamiento de la información a transmitir en un buffer durante un intervalo, enviándola posteriormente a la red a la tasa media obtenida durante dicho intervalo. Sin embargo, estas técnicas de almacenamiento añaden un retardo a la transmisión que no es admisible en el caso de servicios interactivos. En este capítulo se propone y analiza un nuevo método que reduce el retardo introducido por el almacenamiento mediante el uso de técnicas de predicción de tasa binaria. Dicha predicción se lleva a cabo en base a la caracterización del tráfico de vídeo generado por un codificador MPEG mediante un modelo autoregresivo e integrativo de media móvil (ARIMA). La identificación del modelo ARIMA se ha realizado empleando tres secuencias patrón de larga duración (34000, 51000 y 174000 cuadros) codificadas en MPEG-1. Como resultado se ha observado la invarianza temporal de los coeficientes del predictor, junto con su insensibilidad a los cambios en la calidad de imagen seleccionada en el codificador. A partir de estos resultados se propone un nuevo esquema de conformación de tráfico y se analizan las ventajas e inconvenientes respecto a los clásicos sistemas de almacenamiento.

El resto del capítulo está organizado como sigue. En el apartado siguiente se presenta el modelo ARIMA no fraccional para la caracterización de la tasa a nivel de cuadro generada por un codificador MPEG VBR. Este modelo se diferencia del presentado en el capítulo anterior en que su parte integrativa no es fraccional, lo cual tiene dos implicaciones. En primer lugar, el modelo no es válido para la síntesis

artificial de tráfico ya que su media no está acotada. Por otra parte, no captura con precisión la dependencia a largo plazo presente en el tráfico de vídeo. Sin embargo, este modelo va a ser utilizado para llevar a cabo una predicción a corto plazo, para lo cual es más interesante que el modelo ARIMA fraccional al ser aún más sencillo de implementar obteniendo prácticamente los mismos resultados. Así, en la sección 3, se hace uso de dicha caracterización para realizar un predictor de tasa, cuyo correcto funcionamiento es comprobado para las tres secuencias bajo estudio. En el apartado 4, se aborda el problema de la conformación del tráfico comentado, comparando los clásicos sistemas de almacenamiento con un nuevo esquema basado en el predictor. Como se verá, la principal aportación del nuevo esquema es la reducción del retardo introducido por la conformación, dado que no se basa en el almacenamiento de muestras pasadas sino en la predicción de muestras futuras. Además, este retardo puede mantenerse por debajo de una cota establecida a priori. A continuación, se lleva a cabo una comparación entre la cantidad de recursos de red necesaria para una fuente de vídeo cuando ésta entrega el tráfico conformado y sin conformar, para finalizar con el estudio de ciertos casos particulares de multiplexación.

2 Modelo ARIMA no fraccional para la tasa a nivel de cuadro

En este trabajo se ha desarrollado un nuevo predictor de tasa binaria para el tráfico de vídeo VBR MPEG a nivel de cuadro. Este predictor se basa en la caracterización previa de dicho tipo de tráfico como proceso ARIMA no fraccional. Dicho modelo ha sido previamente presentado en [Mat96][CruAli97]. Para su desarrollo, se utilizaron tres largas secuencias de vídeo codificadas en MPEG con los parámetros que se indican en la tabla 4.1.

SECUENCIA	CUADROS	M	N	Q
JURASSIC PARK	175000	2	6	6
AMERICA	34000	2	6	9
GEOGRAFÍA DE CATALUÑA	50000	2	6	6

Tabla 4.1. Secuencias utilizadas para la elaboración del modelo ARIMA no fraccional

Las dos primeras secuencias presentan las características típicas de actividad y complejidad, mientras que la tercera se caracteriza por la complejidad y corta duración de las escenas. Asimismo, se han contrastado los resultados obtenidos con la codificación de la secuencia "América" con parámetros (Q=9,M=2,N=4).

En la figura 4.1 se presenta el esquema de un modelo ARIMA(p,d,q). Para elaborar el modelo buscado para la tasa de vídeo a nivel de cuadro se procede inicialmente a la determinación de la parte integrativa. La dependencia a largo término provoca que la tasa media de grupos de imágenes varíe suavemente. Esta variación llega a alcanzar

niveles máximo y mínimos muy distantes. Sin embargo, la varianza se mantiene casi constante. Esto nos permite concluir que la parte integrativa del modelo debe ser de orden 1. Por tanto:

$$C(z) = 1 - z^{-1} \tag{4.1}$$

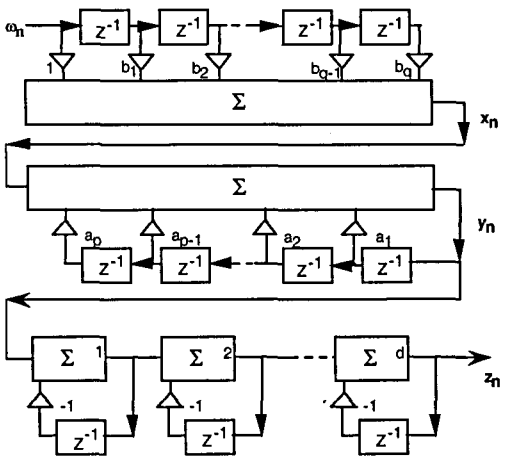

Figura 4.1 Esquema de un filtro ARIMA(p,d,q)

Para determinar las componentes autoregresiva y de media móvil, se emplearon las mismas técnicas que las comentadas en el capítulo 3 para el modelo ARIMA fraccional. Las funciones de transferencia obtenidas son las siguientes:

$$A(z) = 1 - 0.695z^{-6} - 0.3z^{-12} \tag{4.2}$$

$$\begin{aligned} B(Z) = {} & 1 - 0.7618z^{-1} + 0.1136z^{-2} - 0.1676z^{-3} + 0.0195z^{-4} - \\ & - 0.0451z^{-5} - 0.1691z^{-6} + 0.0386z^{-7} + 0.0397z^{-8} - 0.0268z^{-9} + \\ & + 0.0523z^{-10} - 0.0371z^{-11} - 0.1910z^{-12} + 0.1351z^{-13} \end{aligned} \tag{4.3}$$

Finalmente, el modelo ARIMA tendrá como función de transferencia:

$$H(z) = \frac{B(z)}{A(z)C(z)} \tag{4.4}$$

Para evaluar el comportamiento del modelo sintetizado se ha realizado un análisis de los errores de predicción para todas las secuencias. En la figura 4.2 se presenta la autocorrelación de los residuos y los intervalos de confianza para el 99%.

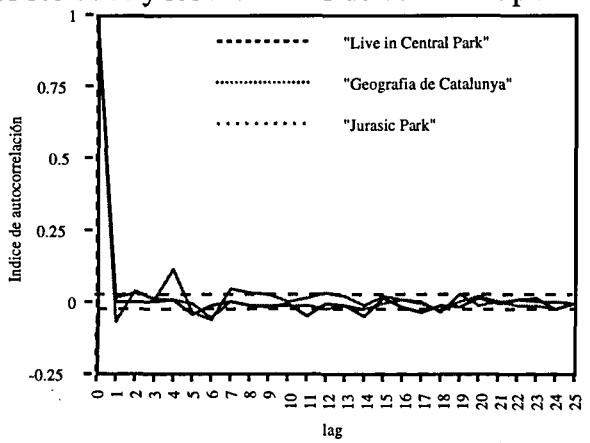

Figura 4.2. Indice de autocorrelación de las series residuales

3 Predicción de la tasa de salida del codificador MPEG VBR

A partir de la caracterización obtenida en el apartado anterior para la tasa generada por cuadro por el codificador de vídeo, en esta sección se plantea la posibilidad de llevar a cabo una predicción de dicha tasa en función de sus valores anteriores. Como ya se ha comentado, dicha predicción será utilizada en la conformación del tráfico previa a su entrega a la red.

Recordando que la componente integrativa del modelo obtenido es de orden 1, podemos expresar dicho modelo de la siguiente manera:

$$y(n) = b_0 w(n) + b_1 w(n-1) + \cdots + b_q w(n-q) + \\ + a_1' y(n-1) + a_2' y(n-2) + \cdots + a_{p+1}' y(n-p-1) \tag{4.5}$$

donde los coeficientes a_i' se obtienen de la convolución de los anteriores coeficientes a_i y c_i.

Partiendo de (4.5), la predicción de la muestra $n+1$ a partir de la muestra n y anteriores sería de la forma:

$$\hat{y}(n+1) = b_0 \hat{w}(n+1) + b_1 w(n) + \cdots + b_q w(n-q+1) + \\ + a_1' y(n) + a_2' y(n-1) + \cdots + a_{p+1}' y(n-p) \tag{4.6}$$

Sin embargo, los valores de la serie $w(n)$ en este contexto de predicción son desconocidos, ya que el predictor trabajará únicamente con los valores anteriores de la serie $y(n)$. Además, el valor $\hat{w}(n+1)$ es un valor futuro. La mejor predicción que se puede hacer para él es tomar el valor medio de la serie residual, el cual según lo visto en el apartado anterior es igual a cero. Por otra parte, recordando que:

$$y(n) = b_0 w(n) + b_1 w(n-1) + \cdots + b_q w(n-q) + \\ + a_1' y(n-1) + a_2' y(n-2) + \cdots + a_{p+1}' y(n-p-1) \tag{4.7}$$

y que, por tanto:

$$\hat{y}(n) = b_0 \hat{w}(n) + b_1 w(n-1) + \cdots + b_q w(n-q) + \\ + a_1' y(n-1) + a_2' y(n-2) + \cdots + a_{p+1}' y(n-p-1) \tag{4.8}$$

se tiene que, restando (4.8) de (4.7):

$$y(n) - \hat{y}(n) = b_0 \left(w(n) - \hat{w}(n) \right) \tag{4.9}$$

Tomando, al igual que en el caso anterior, la esperanza de $w(n)$ como la mejor predicción posible para $\hat{w}(n)$, y recordando que esta es igual a cero, se obtiene:

$$y(n) - \hat{y}(n) = b_0 w(n) \tag{4.10}$$

de donde:

$$w(n) = \frac{y(n) - \hat{y}(n)}{b_0} \tag{4.11}$$

De esta forma, se obtiene la siguiente predicción, en función de las muestras reales y predecidas anteriores:

$$\hat{y}(n+1)=\frac{b_1}{b_0}\left(y(n)-\hat{y}(n)\right)+\cdots+\frac{b_q}{b_0}\left(y(n-q+1)-\hat{y}(n-q+1)\right)+$$
$$+a'_1 y(n)+a'_2 y(n-1)+\cdots+a'_{p+1} y(n-p) \qquad (4.12)$$

Dicha predicción es representable gráficamente como se observa en la figura 4.3. De esta forma, el predictor nos da la mejor estimación posible para la muestra *n+1* en función de las *n* anteriores. También es posible, utilizando dichas *n* muestras, la estimación de las salidas *n+2*, *n+3*, etc. Es decir, podemos utilizar el mismo predictor y las mismas *n* muestras para la obtención de las predicciones "1 adelante", "2 adelante", etc. Para la obtención de la predicción de la muestra *n+j* no hay más que inyectar al predictor la estimación de la muestra *n+j-1*.

La validez del predictor ha sido comprobada para las tres secuencias de estudio. En la figura 4.4 se muestra la estimación de una serie de 40 cuadros de la película "Jurassic Park". Estos 40 cuadros pertenecen a un cambio brusco de escena en la película, es decir, a uno de los momentos en los cuales la predicción se aleja más de la serie original. El principio corresponde a una escena con poco movimiento, obteniéndose una tasa media de aproximadamente 50000 bits por cuadro. Tras el salto, se pasa a una escena más compleja espacialmente y con mucho movimiento, lo cual provoca tasas alrededor de los 200000 bits por cuadro. Se puede observar como la predicción es muy precisa durante lo que podríamos llamar régimen permanente, y como se adapta rápidamente a los cambios de escena.

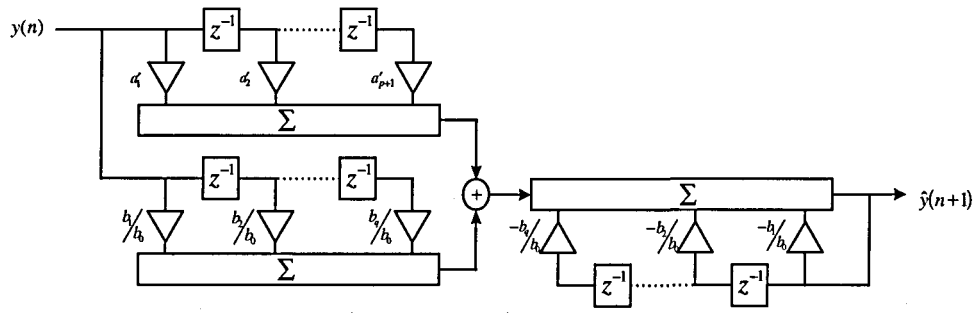

Figura 4.3. Predictor ARIMA

Como se ha mencionado anteriormente, el predictor se puede utilizar también para la obtención de muestras "j adelante". En la figura 4.5 se muestran, para el mismo intervalo de la serie anterior, las predicciones "2 adelante" y "3 adelante". Como se verá en el próximo apartado, estas predicciones van a resultar fundamentales en el buen funcionamiento del conformador de tráfico presentado en este trabajo.

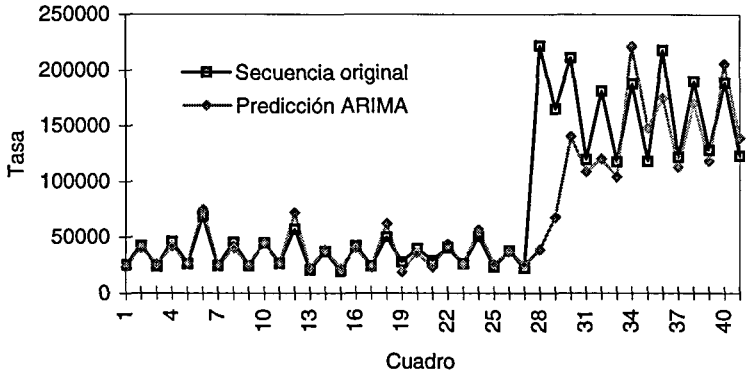

Figura 4.4. Predicción "1 adelante" en cambio de escena

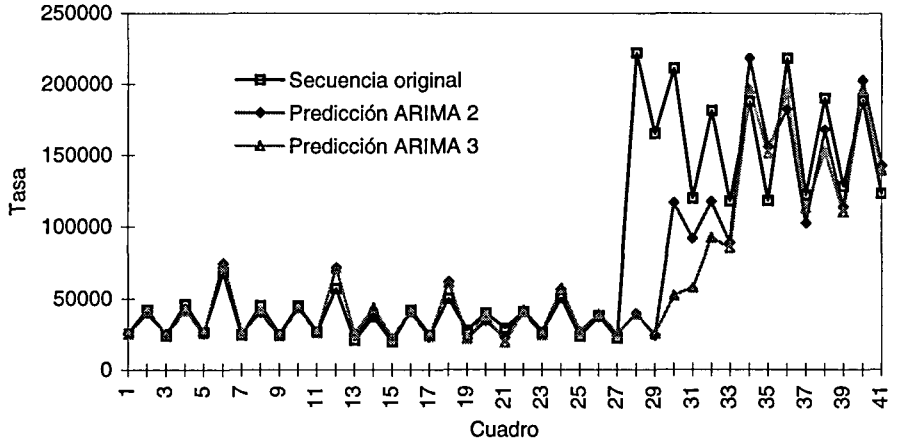

Figura 4.5. Predicción "2 adelante" y "3 adelante"

4 Conformación del tráfico MPEG VBR

Para minimizar la variabilidad de la tasa generada es aconsejable la conformación del tráfico generado por el codificador. Resultados previos que ponen de manifiesto la bondad de estos sistemas se pueden encontrar en [RosFra94][Gra97][ZhaKur97]. Por otra parte, cuando se desea multiplexar distintas fuentes de vídeo sobre un mismo canal, existe el problema de la posible alineación de los GoPs de las secuencias a multiplexar. En el peor de los casos, los cuadros I de las secuencias multiplexadas se entregarán a la vez al multiplexor, con lo que la probabilidad de pérdida de celdas pertenecientes a estos cuadros se verá notablemente aumentada. Recordemos que los cuadros I necesitan una mayor cantidad de bits para ser codificados que los cuadros B y P al no utilizar la compensación de movimiento. Para evitar este problema, aparece de nuevo la solución del suavizado del tráfico previo a la multiplexación.

El suavizado se puede llevar a cabo a través del almacenamiento de las imágenes y su posterior transmisión, una vez determinada la tasa media requerida. Este mecanismo sólo puede ser empleado en servicios que puedan aceptar un retardo de transmisión superior al tiempo necesario para almacenar el grupo de imágenes a suavizar, como por

ejemplo los servicios de difusión de vídeo o de VoD. Para los servicios que no admitan elevados retardos, como la videoconferencia o la videotelefonía, es aconsejable el empleo de técnicas de predicción que permitan, sin introducir retardos adicionales, transmitir a una tasa estimada. De esta manera, se puede reducir la variabilidad del tráfico VBR MPEG de forma similar al empleo de la suavización, sin necesidad de introducir un retardo adicional. Sin embargo, el empleo de técnicas de predicción puede provocar un error de estimación que incremente o decremente la tasa de transmisión de las imágenes notablemente. Este efecto aparecerá cuando se produzca un cambio de escena o en escenas con movimientos de cámara.

En este apartado se presenta un nuevo esquema de conformación de tráfico MPEG VBR, basado en el predictor desarrollado en la sección anterior. Además, sus prestaciones son comparadas con los clásicos sistemas de almacenamiento y con otros de más reciente aparición.

4.1 MÉTODOS CLÁSICOS DE SUAVIZADO

Uno de los sistemas más clásicos de suavizado se basa en el almacenamiento de un número determinado de cuadros, entregándolos después a la red a la tasa media obtenida para todo el grupo. Este tipo de suavizado se ha llamado en otros trabajos suavizado ideal [LamCho96], ya que calcula exactamente la tasa media a la que debe entregar la información a la red. Generalmente, el número de cuadros utilizado para este suavizado es N, es decir, el número de cuadros de un GoP. Así, llamando $S(n)$ al número de bits necesarios para codificar el cuadro n, la tasa entregada a la red durante el siguiente GoP será:

$$r = \frac{S(n) + S(n+1) + \cdots S(n+N-1)}{N\tau} \tag{4.13}$$

donde hemos denominado con τ al tiempo de cuadro.

Por la propia filosofía de este tipo de suavizado, un cuadro en concreto puede sufrir un retardo de almacenamiento previo antes de ser entregado a la red de hasta $2N\tau$ segundos. En otras palabras, en un sistema funcionando a 25 cuadros por segundo y con 6 cuadros por GoP, una imagen se puede ver retardada hasta 480 ms antes de ser entregada a la red. Este retardo, si bien puede ser admisible para un servicio de difusión, no lo es para servicios interactivos. En la figura 4.6 se muestra una porción de la secuencia "Jurassic Park", y su conformación mediante el suavizado ideal. En ella se observa como durante cada GoP se transmite a la tasa media del GoP anterior. Por otra parte, el retardo sufrido por los cuadros representados en la figura anterior se puede observar en la figura 4.7. La unidad de tiempo adoptada es el tiempo de cuadro. Recuérdese que dicha secuencia ha sido codificada con un parámetro N igual a 6.

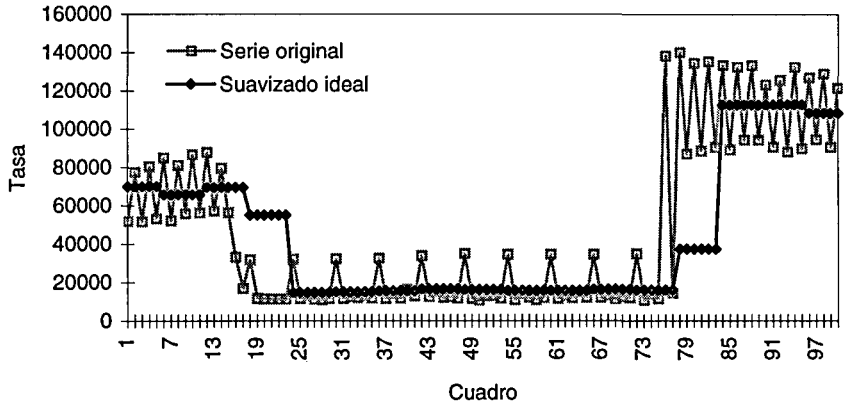

Figura 4.6. Suavizado ideal

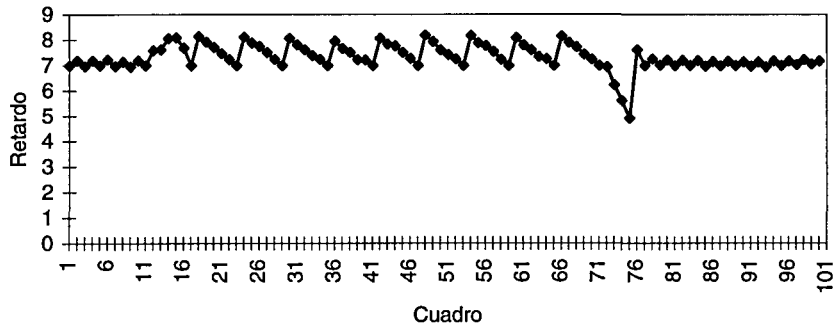

Figura 4.7. Retardo por cuadro con suavizado ideal

Otro tipo de conformación posible consiste en tomar como tasa de salida de un cuadro el promedio de los *N* cuadros anteriores, e ir actualizando dicha tasa para cada cuadro. El suavizado llevado a cabo por este método, que podemos llamar de ventana deslizante, se muestra en la figura 4.8, y su retardo en la figura 4.9.

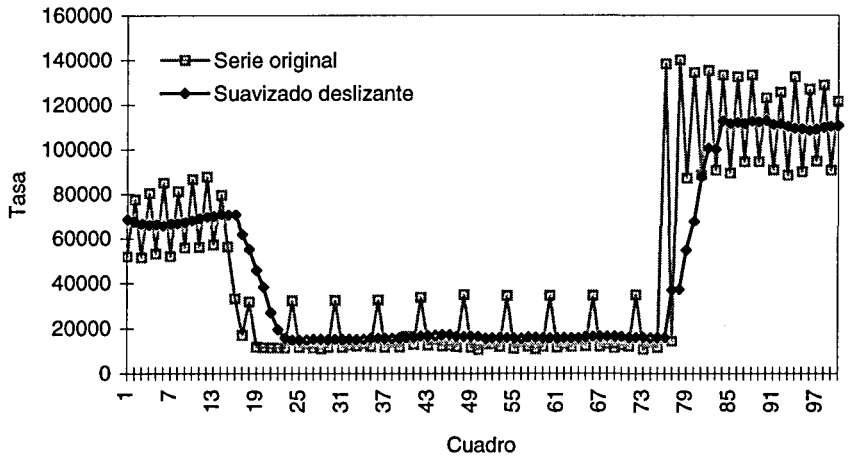

Figura 4.8. Suavizado deslizante

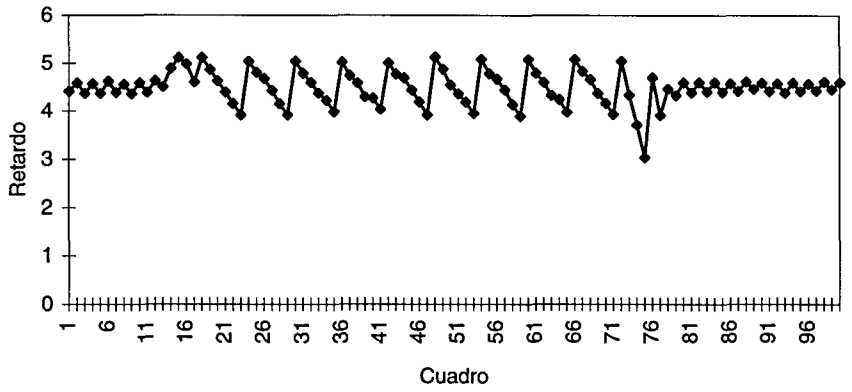

Figura 4.9. Retardo por cuadro con suavizado deslizante

Los resultados obtenidos para las tres secuencias bajo estudio y para cada uno de los dos modelos de suavizado anteriores se presentan en las tablas 4.2 y 4.3. En la primera de ellas se presentan los parámetros estadísticos referidos a la tasa de bits por cuadro, mientras que en la segunda se analizan los valores de los retardos experimentados por cada cuadro. Se puede observar como en cada caso se ha reducido de forma considerable tanto el coeficiente cuadrático de variación de la serie de salida, C_r^2, como su relación de rafagueo, B_r. Dichos parámetros se definen de la siguiente forma:

$$C_r^2 = \left(\frac{\sigma_r}{\bar{r}}\right)^2 \qquad B_r = \frac{r_{pico}}{\bar{r}} \tag{4.14}$$

donde r representa la tasa media de salida del conformador, r_{pico} la tasa máxima y σ_r la desviación estándar. Con ambos parámetros se obtiene una idea de la variabilidad del tráfico respecto a su media.

En el peor de los casos, la reducción de la relación de rafagueo es del 20%, lo cual incidirá en una considerable mejora a la hora de asignar recursos en una red ATM.

SECUENCIA	SUAVIZADO	MIN	MAX	MEDIA	σ	C_r^2	B_r
América	Original	5856	92942	25537	14818.5	0.34	3.64
	Ideal	0	63769	25537	7763.13	0.09	2.50
	Deslizante	0	63764	25537	7749.63	0.09	2.50
Geografía de Cataluña	Original	4383	297635	59672	46919.5	0.62	4.99
	Ideal	0	240943	59672	38433.9	0.41	4.04
	Deslizante	0	240939	59672	38448.9	0.42	4.04
Jurassic Park	Original	7307	261901	43924	31556.8	0.52	5.96
	Ideal	0	207708	43924	23164.6	0.27	4.73
	Deslizante	0	209392	43924	23155.8	0.28	4.77

Tabla 4.2. Tasas de bit por cuadro para sistemas de suavizado con almacenamiento

SECUENCIA	SUAVIZADO	MIN	MAX	MEDIA
América	Original	1	1	1
	Ideal	4.4	7.6	6.4
	Deslizante	3.6	6.1	4.8
Geografía de Cataluña	Original	1	1	1
	Ideal	3.5	8.9	6.5
	Deslizante	3.1	6.9	4.8
Jurassic Park	Original	1	1	1
	Ideal	4.5	9.6	7.4
	Deslizante	3.0	10.2	5.6

Tabla 4.3. Retardos para sistemas de suavizado con almacenamiento

El principal inconveniente de ambos métodos de suavizado es, como se ha comentado anteriormente, el retardo introducido a causa del almacenamiento, que llega a ser de más de 10 tiempos de cuadro en el peor de los casos dentro de las simulaciones estudiadas. Hay que tener en cuenta que la calidad del servicio ofrecida vendrá determinada por el retardo máximo que pueda sufrir un cuadro en concreto de la secuencia que se está transmitiendo. Además, conviene recordar que este retardo máximo se acerca al valor de $2N$. En las secuencias utilizadas el valor de N es igual a 6, pero no hay que olvidar que a menudo se utilizan valores superiores para este parámetro. Por ejemplo, un valor muy utilizado para N es el de 12 cuadros, el cual podría llevarnos a retardos de hasta 960 ms.

Para concluir el estudio de la bondad de los sistemas de suavizado clásicos, se puede proceder al estudio de la función de autocorrelación. Esta función presenta picos periódicos cuando se calcula sobre las series sin suavizar, debido al perjudicial comportamiento periódico presentado por estas. Es decir, el sistema de suavizado será tanto más bueno cuanto más extraiga dicha periodicidad en la función de autocorrelación. En las figuras 4.10 y 4.11 se representa la función de autocorrelación de la secuencia "Jurassic Park", junto con la de las salidas suavizadas. En ellas se destaca claramente como se han eliminado los picos periódicos de tasa elevada que tanto perjudican la eficiente asignación de recursos.

En definitiva, se puede concluir que para ambos métodos las prestaciones ofrecidas en cuanto a conformación de tráfico son buenas para servicios que no presenten fuertes requisitos de retardo. Sin embargo, si dichos requisitos son más restrictivos, los sistemas de almacenamiento pueden hacer bajar la calidad del servicio por debajo de la mínima deseada.

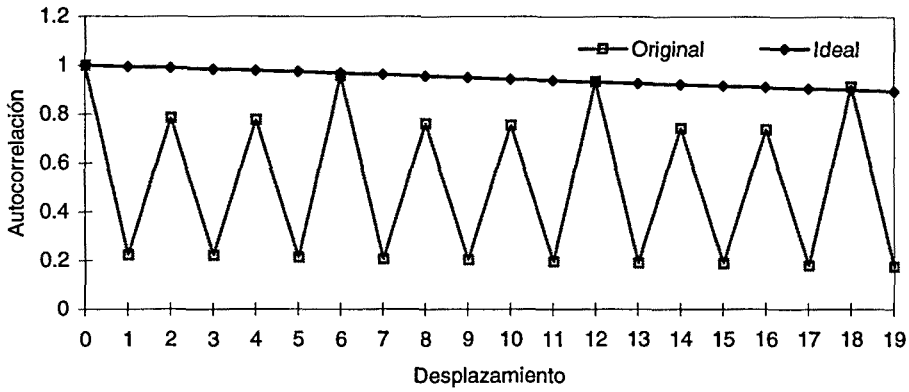

Figura 4.10. Autocorrelación de la serie conformada mediante el suavizado ideal

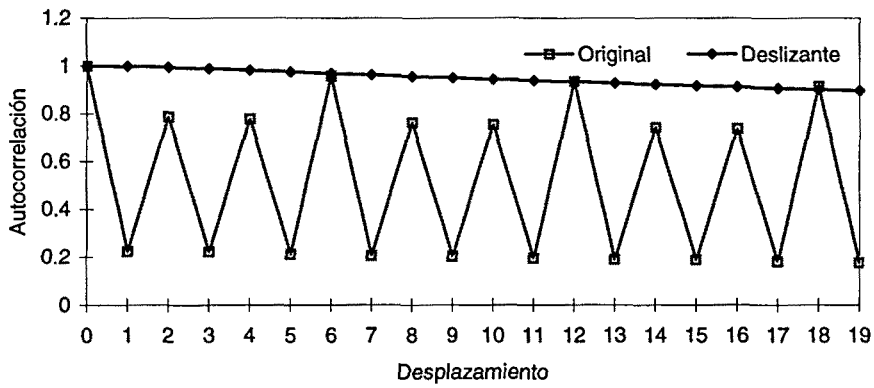

Figura 4.11. Autocorrelación de la serie conformada mediante el suavizado deslizante

4.2 SUAVIZADO DE LAM

Un método nuevo de suavizado muy reciente es el presentado por Lam y otros en [LamCho96]. Se trata de un método predictivo, que acota el retardo máximo introducido y que trata de minimizar en lo posible la variación de la tasa de salida del conformador. Así, presenta una principal ventaja respecto a los métodos anteriores en cuanto a que consigue mantener el retardo por debajo de un valor predeterminado.

Sin embargo, la parte predictiva del método no es muy precisa. Se basa simplemente en tomar como valor de tasa para los cuadros futuros el valor de los cuadros anteriores que ocupaban la misma posición dentro del GoP. Esta predicción puede ser correcta cuando no se producen cambios importantes dentro de la secuencia de vídeo que se está codificando. Durante esos instantes los distintos cuadros dentro de cada GoP se van codificando aproximadamente con el mismo número de bits, siendo una buena aproximación tomar el valor del cuadro n para predecir el valor del cuadro $n+N$. Esto implica que mientras no haya fuertes transiciones de tasa, el suavizado proprocionado por este método es suficientemente bueno.

El problema aparece cuando se producen variaciones de escena en la secuencia que se está codificando, las cuales dan lugar a variaciones bruscas de la tasa de salida del

codificador MPEG VBR. Estas variaciones van a provocar que la predicción utilizada no sea correcta, y que el fallo se propague durante todo un GoP. El resultado de todo esto es la aparición de picos de tasa de salida muy elevados, que afectan negativamente al suavizado de la secuencia. Además, es posible que estos picos provoquen que no se cumpla el contrato establecido con la red, lo cual conlleva el posible descarte de celdas. En la figura 4.12 se presenta una porción de la secuencia "América" y su conformación mediante este método, donde se observa la aparición de los picos de tasa comentados.

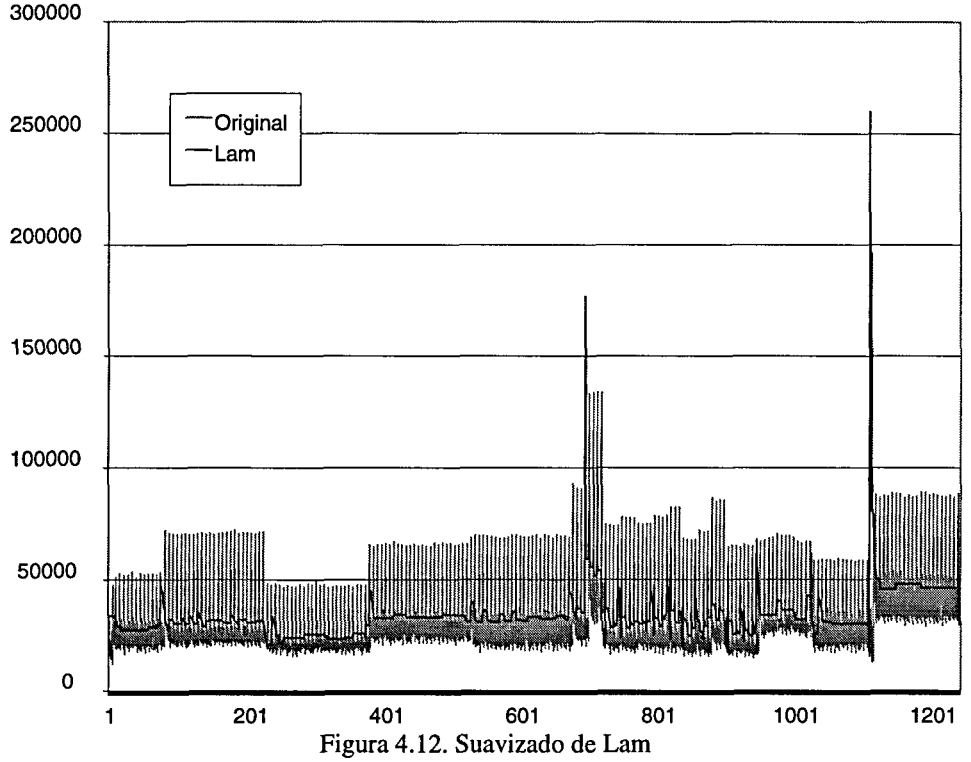

Figura 4.12. Suavizado de Lam

4.3 SUAVIZADO MEDIANTE FILTRADO PASO BAJO

San-Qi Li y otros proponen un método de suavizado basado en un filtrado paso bajo de la señal de vídeo a transmitir [LiCho95]. En realidad, este método es muy parecido al de ventana deslizante presentado anteriormente. Disminuyendo el tamaño de la ventana sobre la cual se hace el promedio para obtener la tasa de salida del conformador se consigue disminuir el retardo introducido, a costa de un empeoramiento en la calidad del suavizado. Por el contrario, si se aumenta el tamaño de la ventana, el suavizado mejora pero aumenta el retardo. Esta variación del tamaño de la ventana se puede entender como la variación en la frecuencia de corte de un filtro FIR con todos sus coeficientes iguales.

El estudio de Li presenta una pequeña mejora ajustando un filtro de Butterworth de orden 2. En la figura 4.13 se presenta este tipo de conformación para distintos valores de la frecuencia de corte del filtro paso bajo empleado y sus valores correspondientes de

retardo máximo *D*. Puede observarse como la calidad del suavizado decrece rápidamente al hacer más restrictiva la condición de retardo máximo.

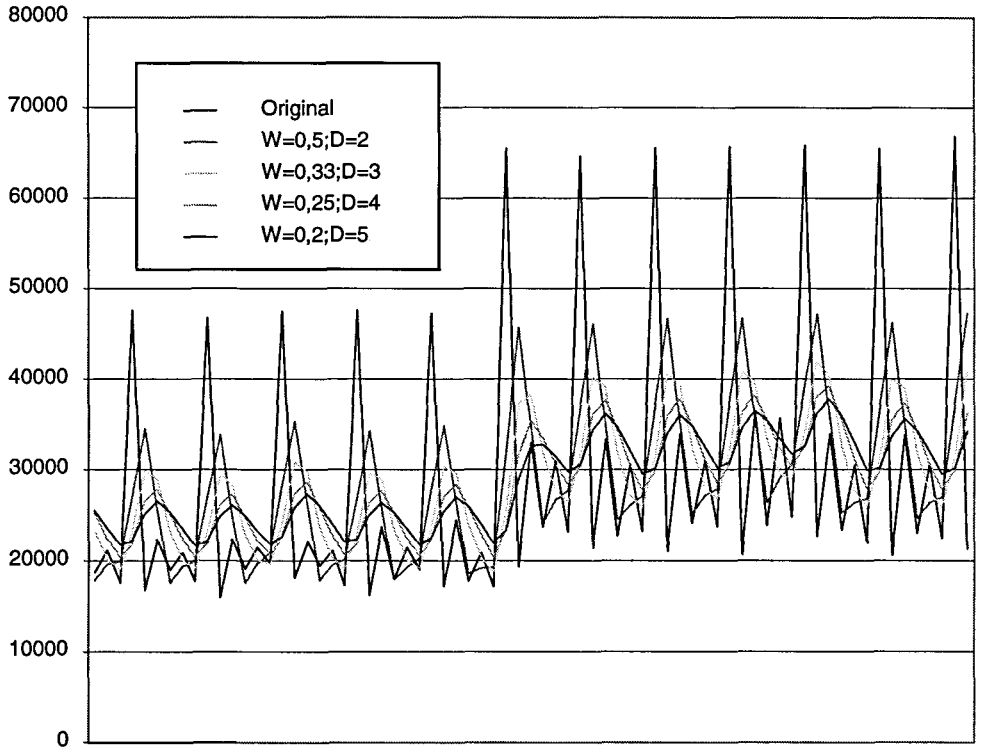

Figura 4.13. Conformación mediante filtrado paso bajo

4.4 CONFORMACIÓN PREDICTIVA

Con objeto de adaptar la conformación del tráfico a servicios con fuertes restricciones temporales, se introduce en este apartado un nuevo método de conformación basado en la predicción de muestras futuras. Parte de los resultados de este trabajo han sido publicados en [CruAli97a] y [CruAli98]. En su forma más general, el conformador de tráfico presenta la estructura de la figura 4.14. En ella se puede observar un buffer de almacenamiento, donde se coloca la información correspondiente a los nuevos cuadros, y que será extraída a la velocidad que le indique el controlador. Dicho controlador, en los métodos de almacenamiento estudiados anteriormente, no haría más que promediar las longitudes de los *N* cuadros anteriores, para obtener la velocidad de extracción del buffer. En el caso del suavizado ideal esta velocidad sería constante durante todo un GoP, mientras que en el caso de ventana deslizante se iría actualizando cuadro a cuadro. Por otra parte, en el método presentado por Li el controlador consistiría en el filtro de Butterworth propuesto.

Para el suavizado predictivo, el controlador estará formado por un predictor ARIMA y dos registros tal y como se muestra en la figura 4.15. La idea consiste en la utilización de un número determinado de muestras pasadas y otro de muestras futuras predecidas para el cálculo de la tasa de salida. Según el dibujo de la figura, se utilizarían

K muestras en total, de las cuales L_1 serían pasadas, incluyendo la actual, y por tanto exactas, y L_2 serían muestras futuras. Es decir, cada vez que la tasa de un cuadro nuevo llega al conformador, se almacena dentro del grupo L_1. Por otra parte, con esta misma tasa como entrada, se hace funcionar al predictor L_2 veces, de forma que se calculan las predicciones "1 adelante", "2 adelante", y así hasta la "L_2 adelante". Promediando esas K muestras, se obtendrá la tasa de salida deseada.

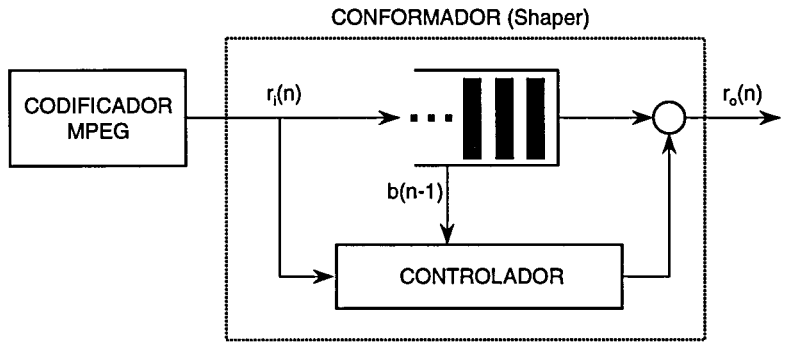

Figura 4.14. Esquema general del conformador de tráfico

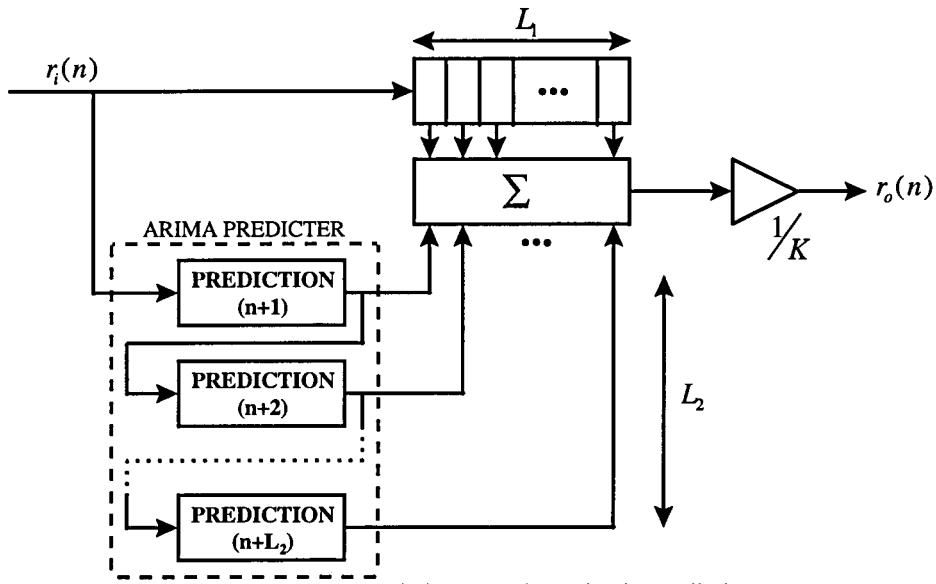

Figura 4.15. Controlador para el suavizado predictivo

Como se comentó anteriormente, uno de los problemas en el suavizado de Lam era la escasa precisión del sistema de predicción utilizado. En la figura 4.16 se compara este sistema de predicción basado en el GoP anterior con el utilizado en este apartado. Como puede observarse, esta última se adapta más rápido que la anterior a los cambios de escena.

En particular, en el conformador presentado en este apartado no se hará necesario tomar un número de muestras pasadas L_1 superior a la unidad, lo cual quiere decir que nos bastará con la muestra actual como única muestra del pasado para calcular la tasa de salida. De esta forma se evita el aumento del retardo. En cuanto al número de muestras predecidas, deberá ser el necesario para completar al menos un GoP. Es decir, como mínimo L_2 deberá ser igual a N-1. Así, se dispondrá de información del tamaño de todos

los cuadros I, B y P dentro del GoP, con lo que la tasa será lo más precisa posible. Por otra parte, la posibilidad de aumentar el número de muestras predecidas para disponer de la información de más de un GoP ha sido rechazado de forma intuitiva en otros trabajos previos [LamCho96]. En este momento es posible corroborar dicha hipótesis, ya que predecir un segundo GoP daría como resultado la misma tasa de salida. Esto es así debido a que la predicción de las tasas del segundo GoP sería prácticamente igual a la de las tasas del primero.

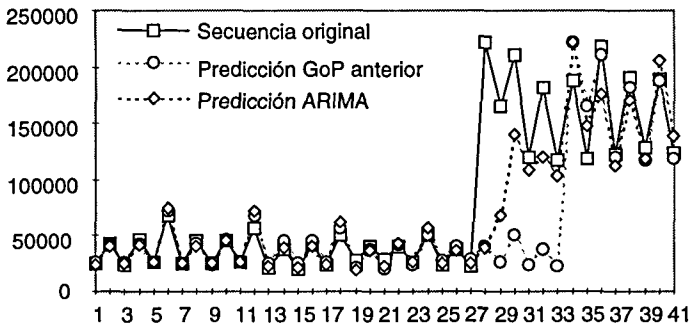

Figura 4.16. Predicción ARIMA y por GoP precedente

Además, se incorpora la posibilidad de utilizar como información adicional el contenido del buffer previo a la introducción de la muestra actual. Esta información será necesaria en el caso de que el conformador quiera garantizar una calidad de servicio determinada, en lo que a retardo de almacenamiento máximo de un cuadro se refiere. El modo de funcionamiento será el mismo que el presentado hasta ahora, con una pequeña modificación que garantizará el cumplimiento de la cota máxima de retardo. Dicha modificación consiste en el cálculo de una tasa mínima de salida para cada cuadro n, en función de su longitud $S(n)$, del contenido del buffer en el momento de su llegada $b(n-1)$, y del retardo máximo permitido, D:

$$r_{min}(n) = \frac{S(n) + b(n-1)}{D} \tag{4.15}$$

El controlador funcionará de la misma forma expuesta anteriormente, pero teniendo en cuenta que existe una cota mínima de salida para cada cuadro que no haya sido aún completamente extraído del buffer. Por tanto, la tasa de salida será como mínimo el máximo de dichas cotas.

El comportamiento del conformador predictivo para las tres secuencias completas bajo estudio se presenta al final del capítulo, en las figuras 4.43 a 4.48. En ellas se puede observar tanto la forma del suavizado como la evolución de los retardos sufridos por cada uno de los cuadros que componen las secuencias. A continuación se lleva a cabo un estudio más detallado, dando más resolución y prestando más interés a las zonas más conflictivas de las secuencias de vídeo, es decir, aquellas zonas en las que se producen cambios bruscos de actividad.

En la figura 4.17 se representa el suavizado obtenido con este tipo de conformador para una fracción de la secuencia "Jurassic Park", en la cual se produce una bajada brusca de tasa. Como se comentará más adelante, estos momentos son los más críticos para este tipo de conformador, y en ellos se activa el mecanismo corrector de tasa introducido anteriormente. Se presentan los casos en los que el retardo está limitado a 2 y 3 tiempos de cuadro, junto con la secuencia original. Es de notar que la secuencia original coincide con la que se tendría si se admitiese un retardo como máximo de 1 tiempo de cuadro. En la figura 4.18 se contemplan los casos de retardo máximo igual a 4, 5 y 6 tiempos de cuadro.

En las figuras se puede observar como al aumentar el máximo retardo permitido se consigue mejor suavizado en las zonas de transición. Los retardos sufridos por los distintos cuadros para cada uno de los casos anteriores se representan en la figura 4.19, donde se puede observar como en ningún caso el retardo supera el máximo permitido.

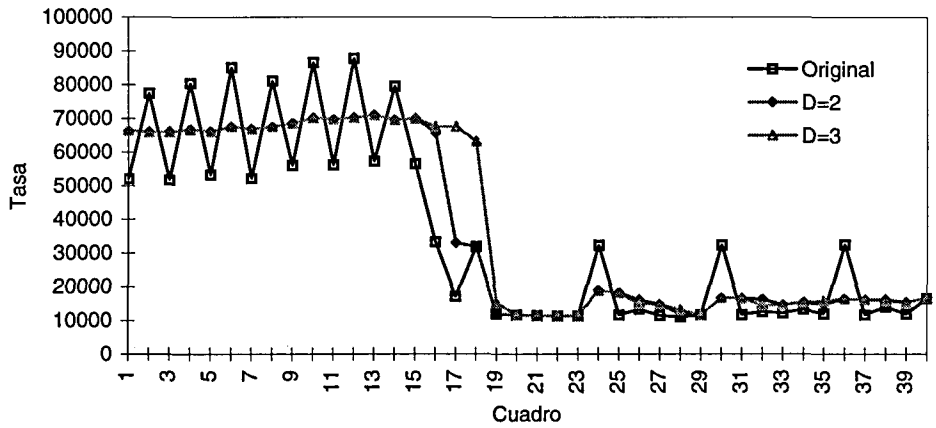

Figura 4.17. Suavizado predictivo en bajada de la serie "Jurassic Park"

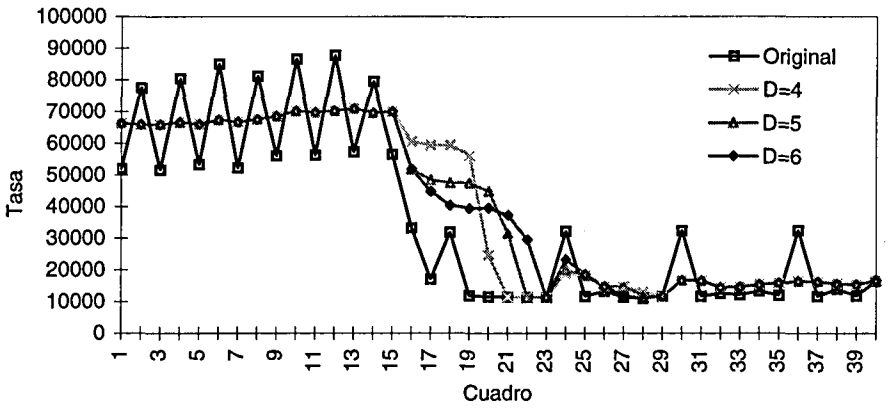

Figura 4.18. Suavizado predictivo en bajada de la serie "Jurassic Park"

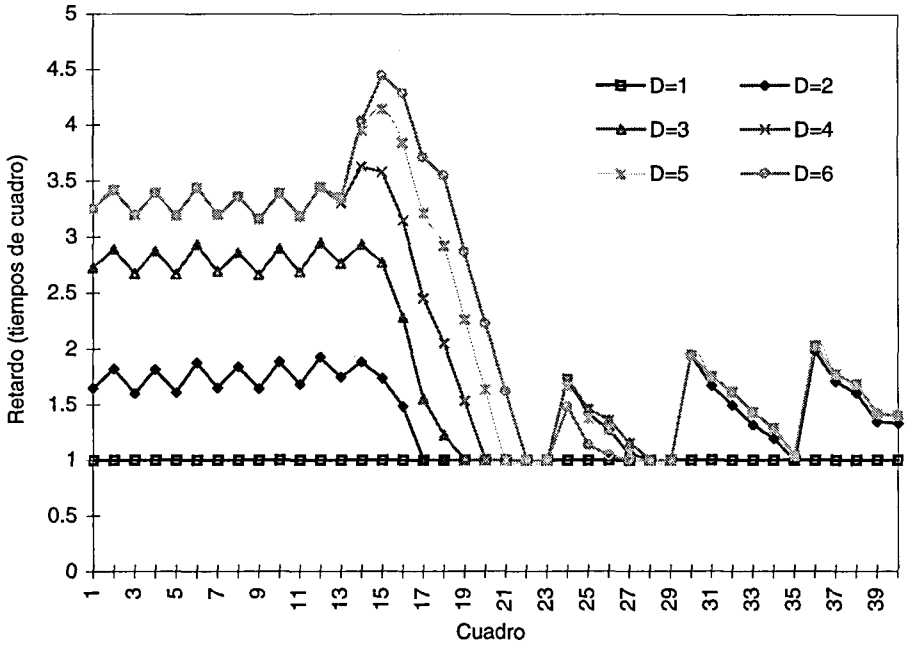

Figura 4.19. Retardo en transición de bajada

El comportamiento del conformador en una transición de subida se muestra en la figura 4.20. En esta ocasión tan sólo se representa hasta el caso D=3, ya que a partir de este valor las curvas son coincidentes con valores de D superiores. El retardo puede observarse en la figura 4.21.

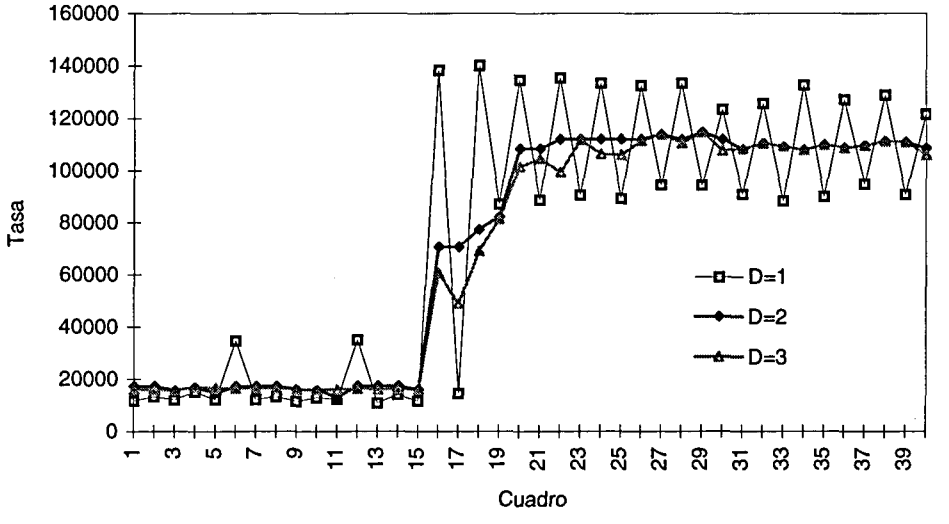

Figura 4.20. Suavizado predictivo en subida de la serie "Jurassic Park"

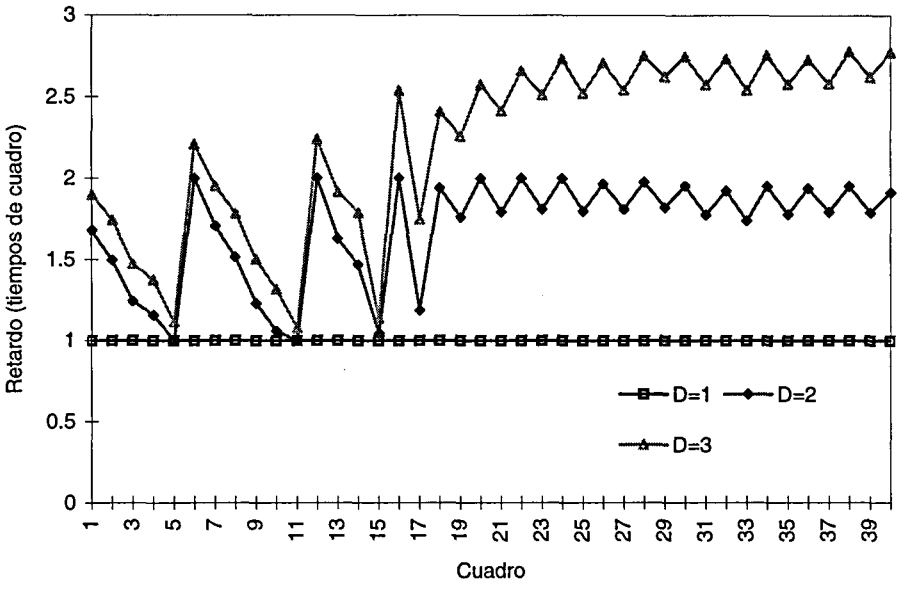

Figura 4.21. Retardo en transición de subida

De las gráficas anteriores se deduce que para este tipo de conformador los momentos más críticos son aquellos en los que se producen bajadas bruscas de tasa. Esto es debido a que el predictor ordenará que las tasas de salida sean bajas, pero en el buffer aún quedará un remanente grande de cuando la tasa era elevada. Es decir, aún quedan cuadros, o parte de ellos, en el buffer, que deben ser extraídos a tasas más elevadas. Es entonces cuando entra en funcionamiento el corrector de tasa, ajustándose a las cotas mínimas previamente calculadas.

En la tabla 4.4 aparecen los resultados estadísticos obtenidos con este conformador, para los casos de estudio, es decir, para retardos máximos de 1, 2, 3, 4, 5 y 6 tiempos de cuadro. Se comprueba como de nuevo se ha conseguido una importante reducción tanto para el coeficiente cuadrático de variación como para la relación de rafagueo. Además, estos valores permanecen prácticamente constantes a partir de D=2.

	BITS / CUADRO						RETARDO		
D	MIN	MAX	MEDIA	σ	C_r^2	B_r	MIN	MAX	MEDIA
1	7307	261901	43925	31516.8	0.52	5.96	1	1	1
2	7741	209530	43925	23691.7	0.29	4.77	1	2	1.4
3	9162	209530	43925	23218.0	0.28	4.77	1	3	2
4	8292	209530	43925	23170.1	0.28	4.77	1	4	2.2
5	8292	209530	43925	23153.3	0.28	4.77	1	5	2.4
6	8292	209530	43925	23145.3	0.28	4.77	1	6	2.6

Tabla 4.4. Conformación predictiva de la secuencia "Jurassic Park"

Con objeto de analizar el margen en el que se mueven los valores de las series original y conformada, en las figuras 4.22 a 4.24 se presentan los histogramas obtenidos

para las tres secuencias bajo estudio. Es interesante comprobar como, al margen de la variación introducida en la función de densidad de probabilidad, patente principalmente en la secuencia "América", se produce una disminución en el margen de valores de las series. Esta disminución es consecuencia directa del suavizado, y es una de las claves a la hora de aprovechar de una forma más eficiente los recursos de red.

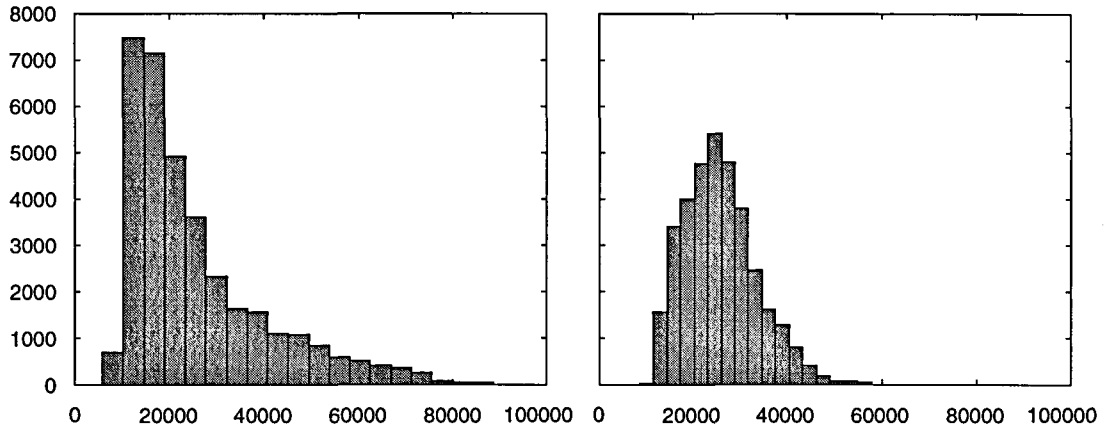

Figura 4.22. Histogramas de las series original y conformada "América"

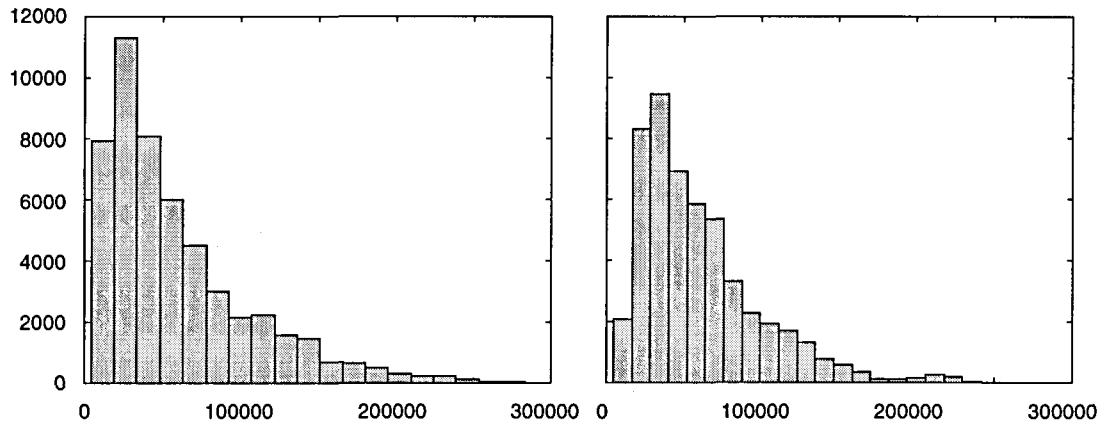

Figura 4.23. Histogramas de las series original y conformada "Geografía de Cataluña"

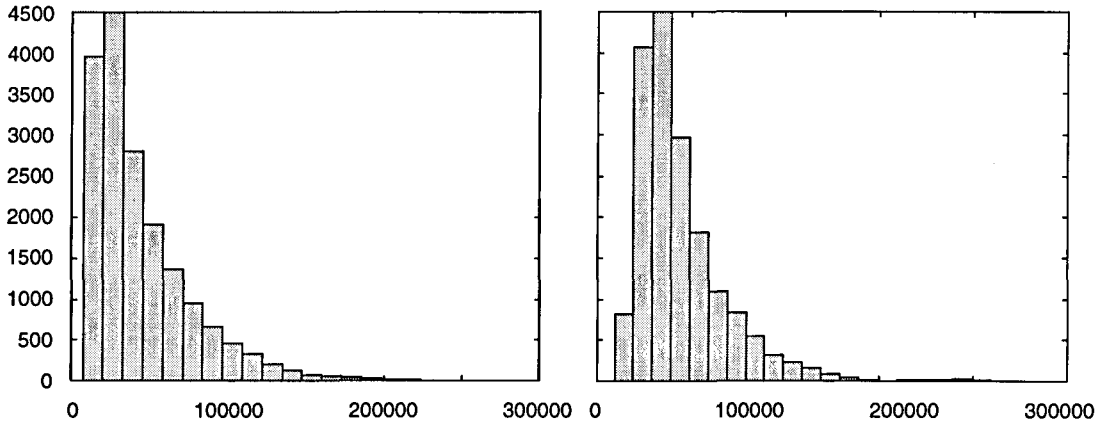

Figura 4.24. Histogramas de las series original y conformada "Jurassic Park"

También se ha estudiado el margen en el que se mueven los valores de los retardos sufridos por cada cuadro individual de cada secuencia. Además, dichos valores se han comparado con los obtenidos utilizando el método de suavizado ideal. Los resultados se

incluyen, en forma de histogramas, en las figuras 4.25 a 4.27. Por una parte, se comprueba de nuevo como el retardo en la conformación predictiva está acotado al valor preestablecido y por debajo del obtenido con el suavizado ideal. Por otro lado, en el conformador predictivo se produce una acumulación de cuadros en los cuales el retardo es igual al valor mínimo de un tiempo de cuadro. Estos valores corresponden en general a los cuadros P y B que son entregados sin problemas en el tiempo de cuadro siguiente al que ellos llegaron. Utilizando el suavizado ideal, sin embargo, la acumulación de cuadros se produce en valores más altos, especialmente en el valor 7 en las secuencias estudiadas. Esto es debido a la necesidad de almacenar un GoP completo, de tamaño 6 en estas secuencias, antes de proceder a la extracción de la información del conformador. La extracción completa se producirá en bastantes ocasiones durante el tiempo de cuadro siguiente, lo que da lugar a la acumulación de valores comentada en el valor 7 tiempos de cuadro. En general, para un GoP de tamaño *N*, la acumulación se presentaría en el valor *N+1*.

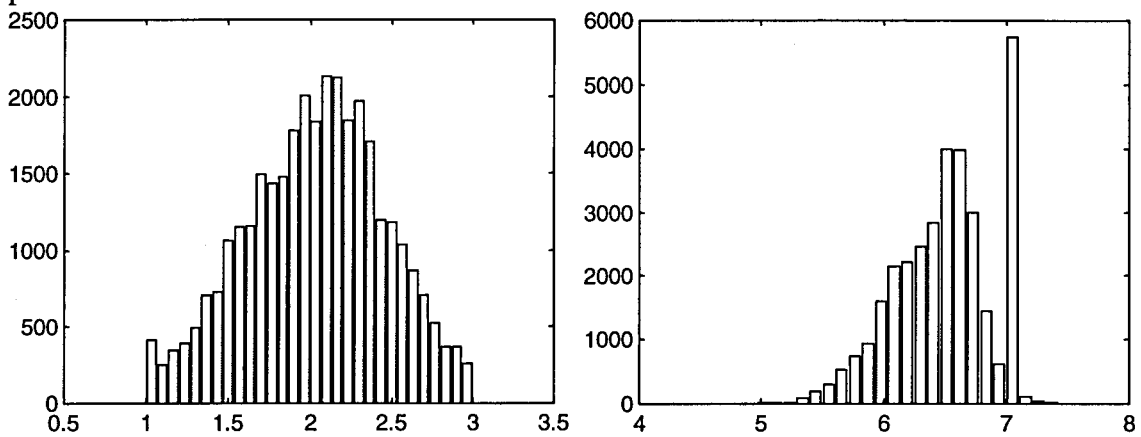

Figura 4.25. Histogramas de retardos en conformación predictiva e ideal. Serie "América"

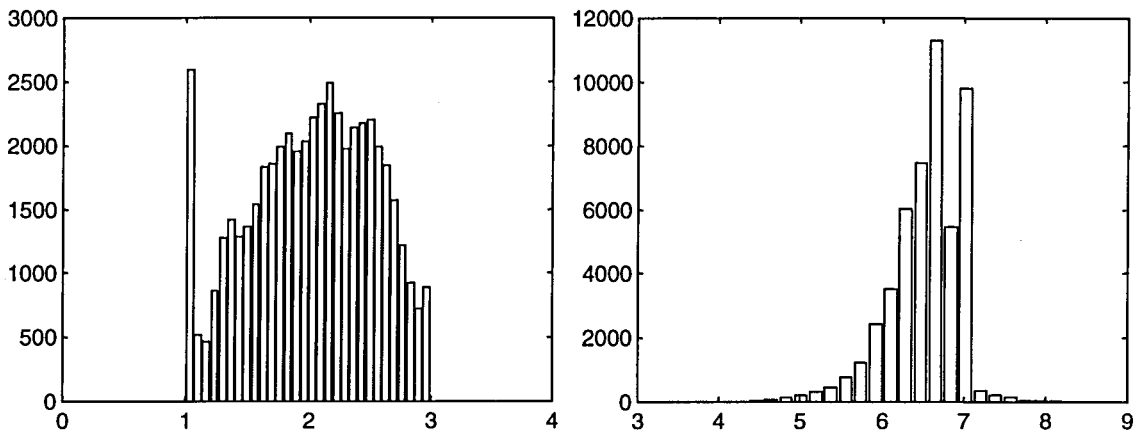

Figura 4.26. Histogramas de retardos en conformación predictiva e ideal. Serie "Geografía de Cataluña"

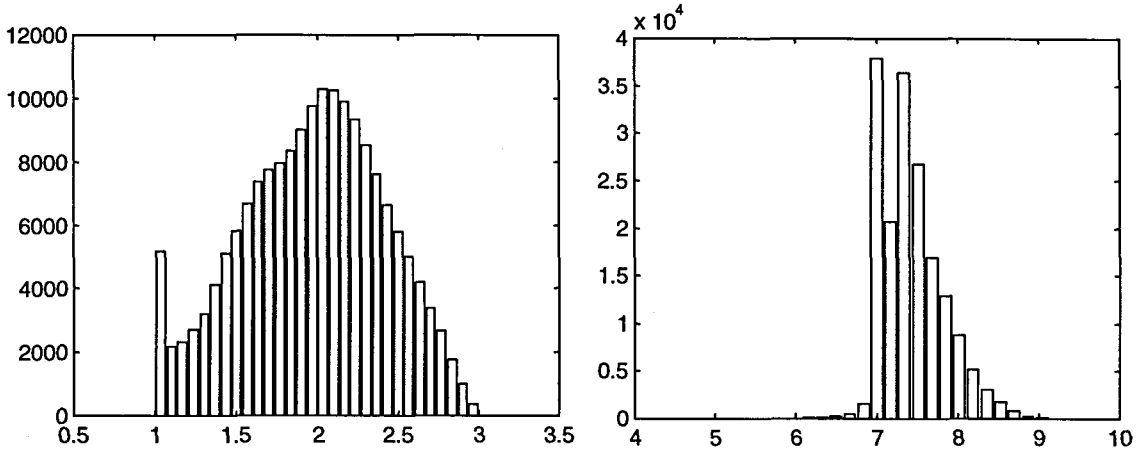

Figura 4.27. Histogramas de retardos en conformación predictiva e ideal. Serie "Jurassic Park"

Para determinar finalmente hasta qué punto es válido el conformador presentado, debemos comprobar la autocorrelación de las series generadas. Recordemos que el principal motivo del suavizado es la eliminación de las llegadas periódicas de tasa elevada. En la figura 4.28 se puede observar dicha función de autocorrelación para los casos D=1, 2 y 3. Como siempre, el caso D=1 (o lo que es lo mismo, la secuencia original), presenta una fuerte correlación. Para D=2, se reduce sustancialmente, pero aún se pueden observar algunos picos cada $kN+1$ desplazamientos. Sin embargo, para D=3 la función queda completamente decorrelada. La diferencia entre estos dos últimos casos se aprecia mejor en la figura 4.29, en la cual se ha suprimido la autocorrelación de la secuencia original con objeto de ganar resolución para los casos D=2 y D=3. Las autocorrelaciones obtenidas para D=4, 5 y 6 ni siquiera se reproducen en la figura, ya que presentan prácticamente la misma forma que la obtenida para D=3. De dichas representaciones se deduce inmediatamente que a partir de una restricción de D=3 conseguimos la máxima decorrelación posible de la serie de salida, lo cual llevaría a una multiplexación estadística óptima de fuentes de este estilo. Si los requisitos de retardo fuesen tan estrictos como para no poder permitir un retardo de suavizado mayor de 2 tiempos de cuadro, es posible la utilización del mismo conformador con un aumento de los recursos necesarios para su ubicación en la red.

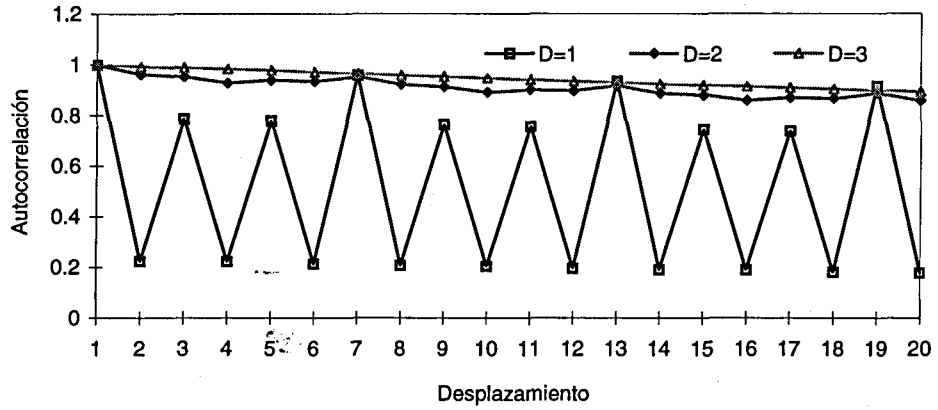

Figura 4.28. Autocorrelación de las series de salida del conformador predictivo

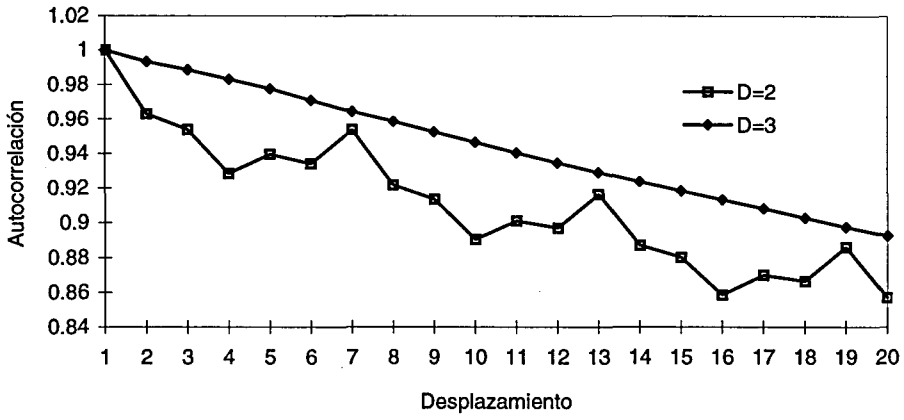

Figura 4.29. Autocorrelación de las series de salida del conformador predictor

Las distintas pruebas presentadas para la secuencia "Jurassic Park", se han realizado también para las otras dos secuencias bajo estudio. Los resultados estadísticos obtenidos para los casos D=3 y D=4 se reflejan en la tabla 4.5.

SECUENCIA	D	BITS / CUADRO						RETARDO		
		MIN	MAX	MEDIA	σ	C_r^2	B_r	MIN	MAX	MEDIO
América	1	5856	92942	25537	14818.5	0.34	3.64	1	1	1
	3	5615	63534	25537	7786.87	0.09	2.49	1	3	2
	4	5615	63534	25537	7765.38	0.09	2.49	1	4	2.6
Geografía de Cataluña	1	4383	297635	59672	46919.5	0.62	4.99	1	1	1
	3	4660	241697	59672	38630.7	0.43	4.05	1	3	1.9
	4	4660	241697	59672	38511	0.42	4.05	1	4	2.3

Tabla 4.5. Resultados estadísticos para las secuencias "América" y "Geografía de Cataluña"

Finalmente, se procede a la comparación de los resultados obtenidos con el nuevo conformador, para el caso en el que el retardo máximo permitido es igual a 3 tiempos de cuadro, con los que se obtuvieron para los sistemas clásicos, utilizando la secuencia "Jurassic Park". En la tabla 4.6 se puede comprobar como la reducción del coeficiente cuadrático de variación y de la relación de rafagueo es aproximadamente igual para todos los casos. Sin embargo, el conformador predictivo proporciona un retardo mucho menor.

SUAVIZADO	BITS / CUADRO						RETARDO		
	MIN	MAX	MED	σ	C_r^2	B_r	MIN	MAX	MED
Ninguno	7307	261901	43924	31556.8	0.52	5.96	1	1	1
Ideal	0	207708	43924	23164.6	0.27	4.73	4.5	9.6	7.4
Deslizante	0	209392	43924	23155.8	0.28	4.77	3.0	10.2	5.6
Predictivo	9162	209530	43925	23218	0.28	4.77	1	3	2

Tabla 4.6. Comparación de los métodos de suavizado

En la figura 4.30 se muestra el suavizado proporcionado por los tres métodos en momentos de gran transición, y en la figura 4.31 se presenta el retardo en dichos momentos. De ambas se deduce el mejor comportamiento presentado por el conformador predictivo.

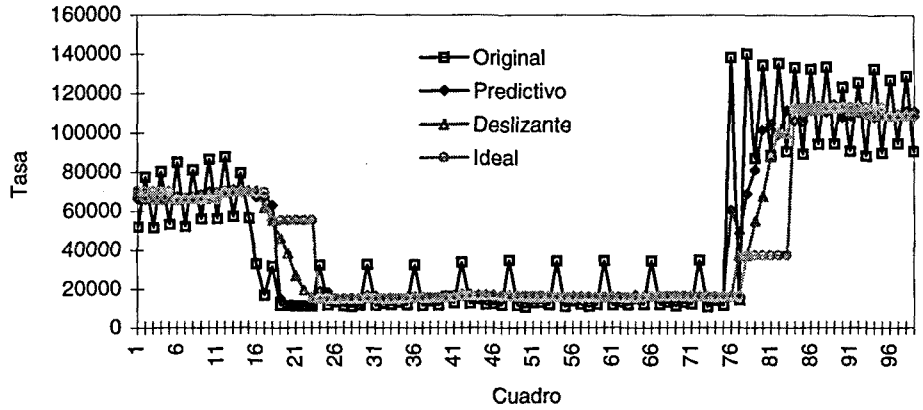

Figura 4.30. Comparación de suavizados

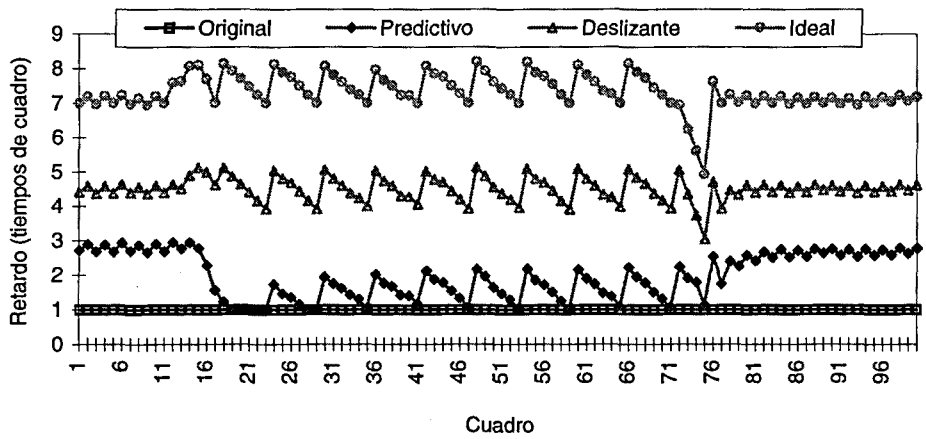

Figura 4.31. Comparación de retardos

5 Transmisión sobre redes ATM

Como se ha venido comentando a lo largo de todo este capítulo, la necesidad de la conformación del tráfico previa a su entrega a la red estriba en la posibilidad de conseguir un mejor aprovechamiento de los recursos de la red. Recordemos que ésta es una de las premisas a la hora del diseño y puesta en funcionamiento de una red de conmutación rápida de paquetes.

5.1 TRANSMISIÓN DE FUENTES SIMPLES

Para comprobar la mejora introducida por el conformador predictivo presentado, se llevaron a cabo una serie de simulaciones utilizando el Simulador Global de redes ATM (SIGLA) desarrollado por los miembros del Grupo de Diseño y Evaluación de Redes y Sistemas de Alta Velocidad del Departamento de Matemática Aplicada y Telemática de

la UPC [CruGar96]. Este simulador permite el análisis y dimensionado de redes ATM desde tres diferentes puntos de vista o niveles: nivel de dispositivo, nivel de red y nivel de servicio. Cada uno de los niveles permite centrarse en aspectos concretos del sistema a estudiar. En concreto, en este estudio se obtuvieron los descriptores de tráfico para la concatenación de las secuencias "Jurassic Park" y "Geografía de Cataluña", tomando como parámetro dos probabilidades de pérdida distintas.

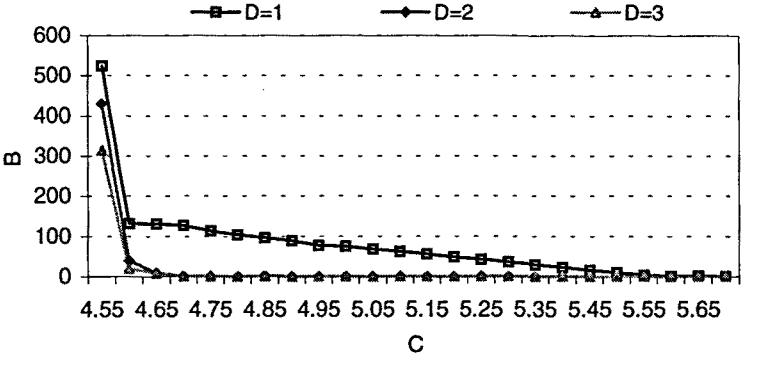

Figura 4.32. Descriptores de tráfico para $P_L = 10^{-4}$

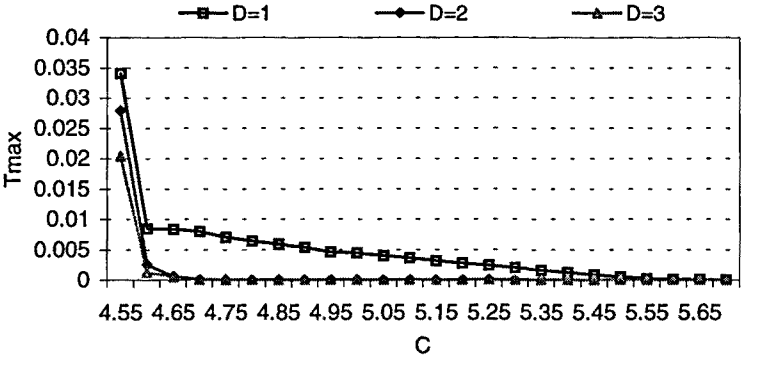

Figura 4.33. Retardo máximo en el multiplexor para $P_L = 10^{-4}$

En la figura 4.32 se representa el tamaño necesario del buffer *B* en función de la capacidad de transmisión asignada *C* para una probabilidad de pérdida de 10^{-4}. El valor de la capacidad está normalizado respecto a la media de toda la secuencia. Los casos que se comparan son los obtenidos con restricciones de retardo en el conformador de 1, 2 y 3 tiempos de cuadro, observándose como la serie original, que coincide con la obtenida para D=1, es la que necesita una mayor cantidad de recursos de red. A la hora de estudiar estas gráficas, debemos centrarnos en el margen de valores que nos ofrezca un retardo en el multiplexor ATM adecuado. Para ello, este retardo se representa para el mismo margen de C en la figura 4.33. Para asegurar valores por debajo de los 10 ms para la fuente sin conformar, debemos tomar capacidades mayores de 4.6 veces la media. Bajar de 3 ms implica una C mínima de 5.2 veces la media. Sin embargo, conformando la fuente el margen de capacidades válidas tiene cotas inferiores mucho más bajas. Por otra parte, asignar una capacidad normalizada igual a la relación de rafagueo equivaldría a estar asignando capacidad a tasa de pico para toda la secuencia,

lo cual evidentemente llevaría a pérdidas nulas sin necesidad de buffer. Este valor de la relación de rafagueo es igual a 6.26 para la secuencia sin conformar y a 5.12 para la secuencia conformada.

Dando mayor resolución a la zona en la cual el retardo de la fuente sin conformar es menor de 10 ms se obtienen las figuras 4.34 y 4.35. En ellas se continúa observando el beneficio del suavizado del tráfico, tanto en recursos de red como en el retardo introducido por el multiplexor ATM. Además, hay que tener en cuenta que todas estas gráficas han sido obtenidas para una fuente simple. En un caso práctico la información generada por esta fuente será multiplexada con la de otras fuentes. En estos casos la mejora será aún mayor debido a la ganancia de multiplexación estadística propia de los conmutadores ATM. Es decir, esta ganancia será mayor cuando se multiplexen fuentes conformadas que si se multiplexan fuentes sin conformar.

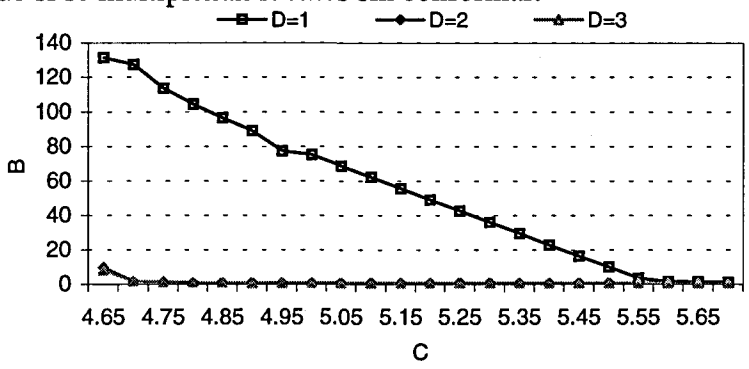

Figura 4.34. Descriptores de tráfico para $P_L = 10^{-4}$

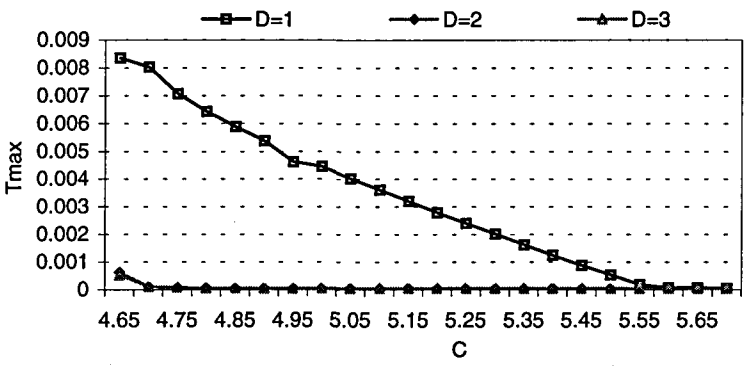

Figura 4.35. Retardo máximo en el multiplexor para $P_L = 10^{-4}$

Otra conclusión que se desprende de las figuras anteriores se refiere al comportamiento cuando la capacidad asignada es muy grande. En esta zona, los beneficios del suavizado no son perceptibles dado que la información apenas es almacenada en el multiplexor ATM cuando ya ha sido posible su retransmisión. Sin embargo, asignar estas altas capacidades a una sola conexión está en desacuerdo con el principio de máxima eficacia en la asignación de recursos.

Con objeto de cuantificar la mejora obtenida mediante la conformación de las fuentes de vídeo MPEG VBR, se puede buscar la reducción en los recursos de red

necesarios manteniendo una calidad de servicio fija. Así, podemos definir la ganancia de suavizado G_s como la relación entre la capacidad de transmisión necesaria para una fuente sin conformar C_{ns} y la de una fuente conformada C_s, manteniendo constante el retardo máximo en el multiplexor ATM:

$$G_s = \frac{C_{ns}}{C_s} \tag{4.16}$$

Tomando un retardo máximo de 1 ms, los valores obtenidos para G_s son de 1.14 y 1.15 para D = 2 y D = 3 respectivamente, lo cual implica unas mejoras del 14% y 15% respectivamente. Aumentando la restricción temporal hasta 200 μs, la mejora supera el 18% en ambos casos. Es conveniente recordar de nuevo que estas mejoras se verán incrementadas al multiplexar fuentes distintas.

En las figuras 4.36 y 4.37 se presentan las curvas obtenidas para una probabilidad de pérdida de 10^{-5}. Esta mayor restricción obliga a que los márgenes de interés comiencen para valores mayores de las capacidades de transmisión, ya que implica buffers mayores que darían lugar a su vez a retardos máximos más elevados.

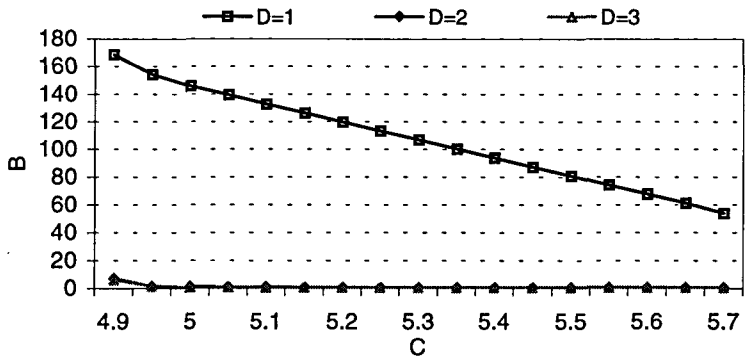

Figura 4.36. Descriptores de tráfico para $P_L = 10^{-5}$

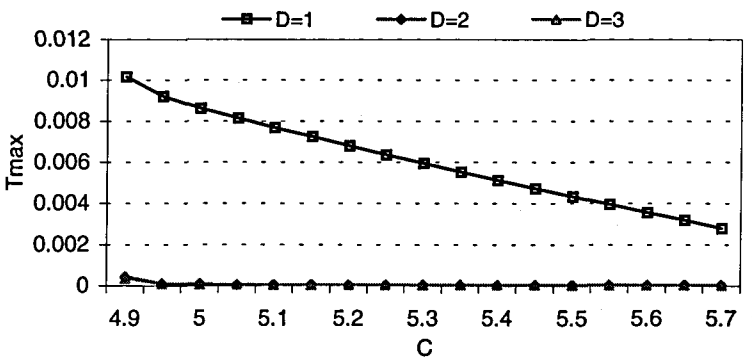

Figura 4.37. Retardo máximo en el multiplexor para $P_L = 10^{-5}$

Otro tipo de simulaciones de interés son las orientadas a la obtención de la relación existente entre la capacidad del enlace y la probabilidad de pérdida obtenida para un tamaño fijo de buffer. Esta relación se muestra en la figura 4.38 para un tamaño máximo de buffer en el multiplexor de 10 celdas ATM. De nuevo es posible observar la mejoría introducida por el suavizado. La secuencia utilizada en este análisis fue la resultante de

la concatenación de las series "Jurassic Park" y "Geografía de Cataluña", con una longitud total de 225000 cuadros, que dieron lugar a 30482000 celdas ATM. Los resultados obtenidos para longitudes de buffer mayores se presentan en las figuras 4.39 y 4.40. En el análisis conjunto de las tres gráficas, se observa como la mejora introducida por el suavizado decrece al aumentar el tamaño del buffer del multiplexor. Este resultado se debe a que este tamaño mayor del buffer es capaz de absorber las ráfagas provocadas por el tráfico sin conformar, con lo que se mantiene una probabilidad de pérdida similar. Sin embargo, hay que recordar que tamaños de buffer grandes implican grandes retardos, lo cual no es admisible por gran parte de los servicios de vídeo.

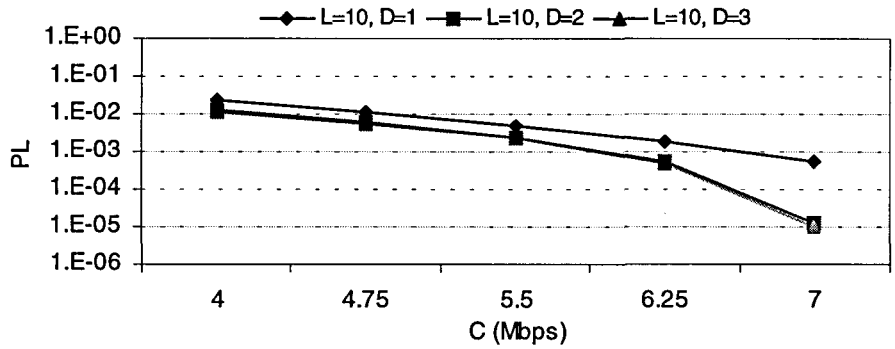

Figura 4.38. Relación entre C y P_L para L=10

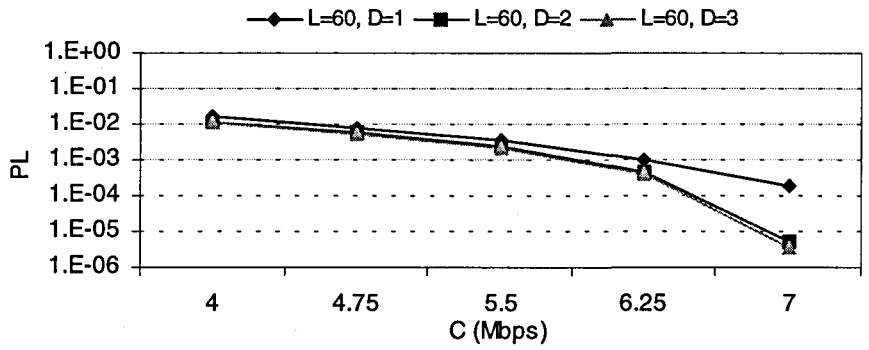

Figura 4.39. Relación entre C y P_L para L=60

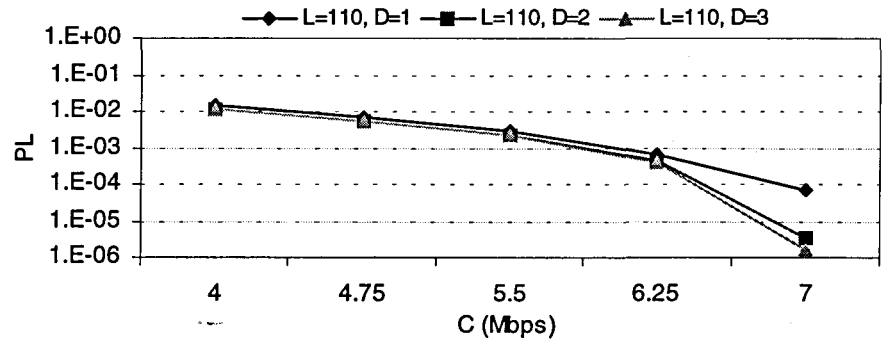

Figura 4.40. Relación entre C y P_L para L=110

5.2 MULTIPLEXACIÓN DE FUENTES DE VÍDEO CONFORMADAS

Al multiplexar varias fuentes sobre un mismo enlace ATM, la ganancia de suavizado se debe hacer de nuevo patente [ZhaKur97]. En particular, puede suceder que varias de las secuencias que se estén transmitiendo lleven sus patrones sincronizados, con lo que los momentos de mayor tasa de una fuente coinciden con los de otras fuentes. Así, es posible estar transmitiendo un cuadro I de una secuencia junto a cuadros I de otras secuencias, lo que generaría altas tasas de transmisión. Además, estas situaciones de alta tasa se irían repitiendo periódicamente cada vez que llegase el cuadro I del nuevo GoP. Existen estudios previos [RosFra94] que apuntan a la necesidad de evitar que diversas secuencias MPEG entren en fase en un enlace ATM. Por una parte, esto llevaría a la necesidad de añadir nuevas formas de señalización en los nodos de acceso, lo cual es bastante complicado en la práctica al implicar un profundo conocimiento del tráfico por parte del nodo. Por otro lado, el problema se hace aún más grave y de difícil solución en los nodos intermedios de la red.

Así, las ventajas de la conformación del tráfico previa a su entrega a la red se hacen de nuevo patentes a la hora de multiplexar fuentes distintas. En la figura 4.41 se muestra como quedaría la transmisión conjunta de dos fragmentos de las secuencias "Jurassic Park" y "América", escogidos al azar, en el caso de que ambas secuencias estuviesen generando sus cuadros I en el mismo momento. Si las secuencias hubiesen sido conformadas previamente, la transmisión quedaría como se muestra en la figura 4.42.

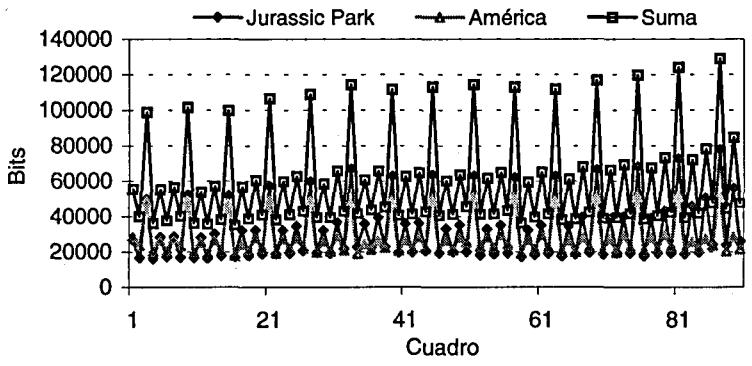

Figura 4.41. Multiplexación de secuencias alineadas sin conformar

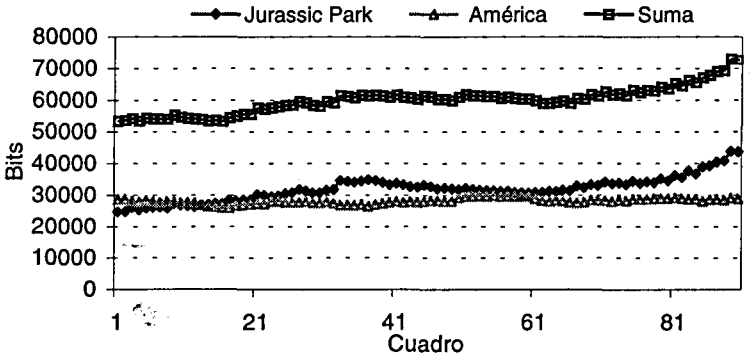

Figura 4.42. Multiplexación de secuencias alineadas previamente conformadas

La ventaja de la conformación se pone de manifiesto estudiando los parámetros que definen la variabilidad de las secuencias. El estudio se ha realizado tanto para el caso de que las secuencias sean transmitidas en fase, como en las figuras anteriores, como para el caso de que las secuencias estén decaladas. Durante la mayor parte de este estudio se ha trabajado con un tamaño de GoP de 6 cuadros y de SGoP de 2 cuadros, lo cual da lugar a una estructura de GoP que sigue el formato IBPBPB. Así, aparte del caso en que los cuadros I de cada secuencia coinciden entre sí, se ha estudiado también qué ocurre cuando el cuadro I de una secuencia coincide con cualquiera de los otros cuadros del GoP de la otra secuencia. Para ello, la secuencia "Jurassic Park" se fue decalando entre 1 y 5 cuadros respecto de la secuencia "América" antes de llevar a cabo la multiplexación. Los resultados obtenidos, tanto para el coeficiente cuadrático de variación como para la relación de rafagueo, se muestran en la tabla 4.7.

De los valores de dicha tabla se pueden extraer una serie de conclusiones. En primer lugar, y como ya se había adelantado, se observa una gran reducción de los parámetros de variabilidad cuando las secuencias son multiplexadas en fase y se suaviza. En particular, el coeficiente cuadrático de variación se reduce en un 95%, mientras que la relación de refagueo decrece un 42%. Recordemos que esta es una situación especialmente peligrosa en caso de que esta coincidencia se produzca cuando se estén multiplexando un número mayor de fuentes. En este caso, es muy posible que el multiplexor ATM no pudiera dar servicio a todas las celdas que fuesen llegando, viéndose en la necesidad de descartar algunas de ellas. Teniendo en cuenta que estas celdas pertenecerían a cuadros I de la secuencia MPEG la situación se agrava, ya que el resto de cuadros del GoP utilizan dicha información como referencia. Como consecuencia, los cuadros afectados serían todos los de la secuencia a transmitir, ya que esta situación se iría repitiendo periódicamente.

	SIN SUAVIZADO		CON SUAVIZADO	
Decalaje en cuadros	C_r^2	B_r	C_r^2	B_r
0	0.19	2.19	0.01	1.27
1	0.05	1.6	0.005	1.2
2	0.13	1.76	0.005	1.2
3	0.05	1.69	0.005	1.2
4	0.13	1.8	0.005	1.2
5	0.05	1.66	0.005	1.2

Tabla 4.7. Comparación de parámetros de variabilidad en secuencias multiplexadas

Por otra parte, el valor de C_r^2 presenta un comportamiento oscilante en la serie agregada conforme se van decalando entre sí las secuencias MPEG. En este caso, el "periodo" es igual a dos, correspondiendo al tamaño del SGoP. El caso en que se

observa la mejoría más pequeña al suavizar es el correspondiente a un decalado de tres tiempos de cuadro. Esta situación era de prever, ya que este decalado corresponde exactamente a la mitad del tamaño del GoP. Decalados superiores hacen que el parecido entre las series vuelva a crecer al ir solapándose un periodo con el inmediatamente anterior o posterior.

Finalmente, es interesante comentar la estabilidad de los valores observados en las distintas secuencias agregadas tras el suavizado, independientemente del decalado entre ellas. Este hecho vuelve a poner de manifiesto la extracción de la correlación de la señal.

Los resultados anteriores, y las grandes mejorías observadas, se deben en parte a que la secuencia de estudio es muy corta, de tan solo 100 cuadros. Muestra de ello es el bajo valor del coeficiente cuadrático de variación en general. Con objeto de obtener valores más fiables, se han repetido los experimentos multiplexando los 175000 cuadros de la serie "Jurassic Park" con ellos mismos, para los casos extremos observados en el estudio anterior. Es decir, se compararán los resultados obtenidos sin decalar las series entre ellas, y decalándolas 3 cuadros. Estos resultados se presentan en la tabla 4.8. Como primera consecuencia válida se puede corroborar el hecho de que los parámetros de variabilidad se mantienen aproximadamente constantes cuando las secuencias que se multiplexan han sido previamente conformadas. Por otra parte, la reducción respecto a la multiplexación sin suavizar continúa siendo importante. En el caso de la relación de rafagueo, el valor se ve reducido entre un 6% y un 13%, mientras que para el coeficiente cuadrático de variación la reducción va de un 13% a un 48%.

	SIN SUAVIZADO		CON SUAVIZADO	
Decalaje en cuadros	C_r^2	B_r	C_r^2	B_r
0	0.5	5.5	0.26	4.77
3	0.3	5.04	0.26	4.74

Tabla 4.8. Parámetros de variabilidad multiplexando la secuencia completa "Jurassic Park"

6 Conclusiones

El tráfico de vídeo MPEG VBR presenta unas fluctuaciones periódicas que deben ser extraídas antes de entregarlo a una red ATM. En este capítulo se ha presentado un nuevo conformador para este tipo de tráfico. Dicho conformador emplea técnicas de predicción, basándose en la caracterización del tráfico real como proceso ARIMA. El modelo sobre el cual se ha desarrollado el predictor es no fraccional, por lo que no es útil en la síntesis de tráfico. Una de las principales características del modelo es la invarianza de los coeficientes obtenida para todas las secuencias codificadas. Es de resaltar que la invarianza se mantiene incluso respecto al cambio en la calidad de la

imagen solicitada al codificador, mediante la variación del parámetro Q. Utilizando como pieza base el predictor ARIMA, se ha propuesto un nuevo conformador de tráfico de vídeo con objeto de mantener la calidad del suavizado y mantener a la vez el retardo adicional introducido por debajo de una cota predeterminada. Así, se pretende que este nuevo sistema sea adecuado para trabajar con servicios de vídeo interactivos, los cuales son muy restrictivos en lo que a retardo de entrega de la información se refiere.

Las prestaciones del nuevo conformador de tráfico han sido comparadas con las de los habituales sistemas de suavizado por almacenamiento, así como con otros sistemas más recientes propuestos en congresos internacionales. Entre los sistemas estudiados, los más clásicos basados en almacenamiento presentan una buena calidad del suavizado, pero introducen un retardo inadmisible en servicios interactivos. Por otra parte, la técnica de filtrado paso bajo permite acotar el retardo, a costa de una disminución en la calidad del suavizado. Finalmente, el método de Lam consigue también acotar el retardo, pero su comportamiento en cambios de escena no es correcto.

El estudio de todos los parámetros comentados se ha llevado a cabo por un lado en base a la representación y estudio temporal de las secuencias bajo estudio. Además, se han analizado los histogramas tanto de las series originales y conformadas, como de los retardos introducidos por los métodos de conformación. Por otro lado, se han obtenido y comparado parámetros estadísticos de primer y segundo orden, así como funciones de autocorrelación.

Finalmente, se han realizado simulaciones para comprobar las mejoras que ofrece el suavizado cuando se transmite el tráfico de vídeo sobre una red ATM, poniéndose de manifiesto de nuevo la necesidad de la conformación. Las simulaciones se han realizado en primer lugar sobre una fuente simple. Por una parte, se han obtenido los descriptores de tráfico fijando el valor de la probabilidad de pérdida. Con objeto de cuantificar la mejora, se ha propuesto el parámetro *Ganancia de Suavizado*, como la relación entre la capacidad de canal necesaria por el tráfico sin conformar y el conformado, para un mismo valor del retardo máximo en el multiplexor ATM. Por otro lado, se ha obtenido la probabilidad de pérdida en función de la capacidad del enlace, para diversos valores del tamaño del buffer del multiplexor, poniéndose de manifiesto de nuevo las ventajas del suavizado para retardos inferiores a 100 ms. Además, se ha estudiado la mejoría obtenida cuando se multiplexan distintas fuentes de vídeo sobre un mismo enlace.

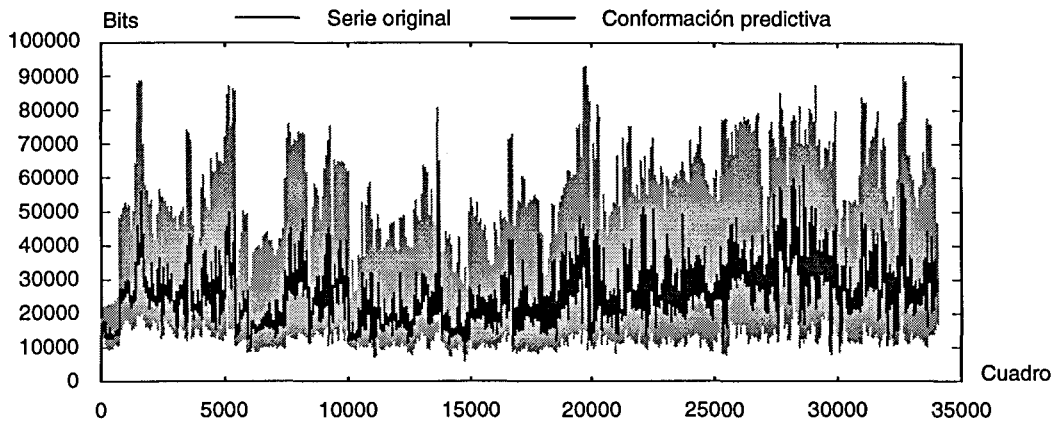

Figura 4.43. Conformación predictiva de la serie completa "América"

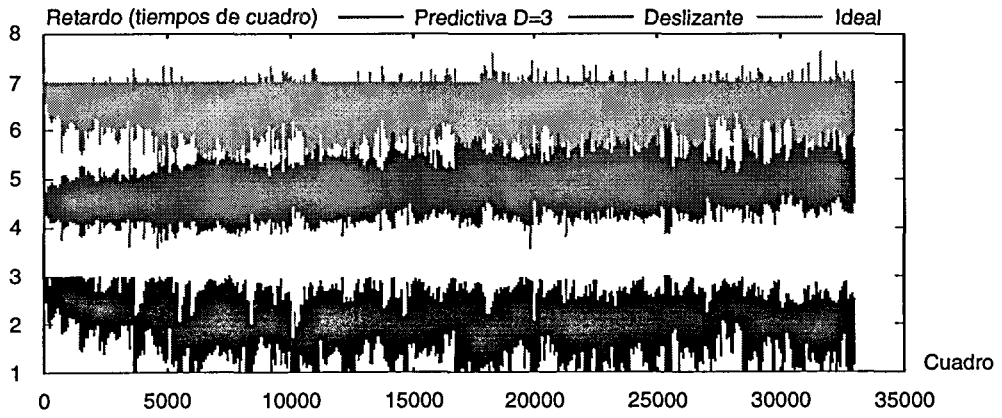

Figura 4.44. Comparación de retardos para la serie completa "América"

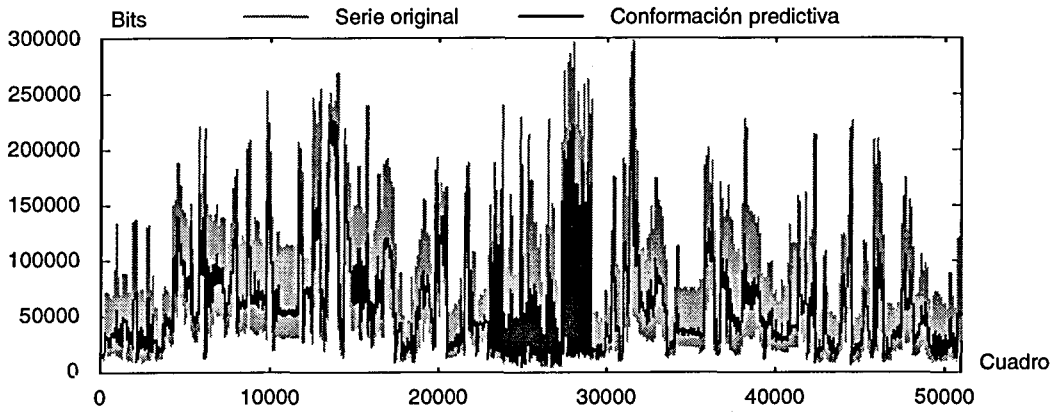

Figura 4.45. Conformación predictiva de la serie completa "Geografía de Cataluña"

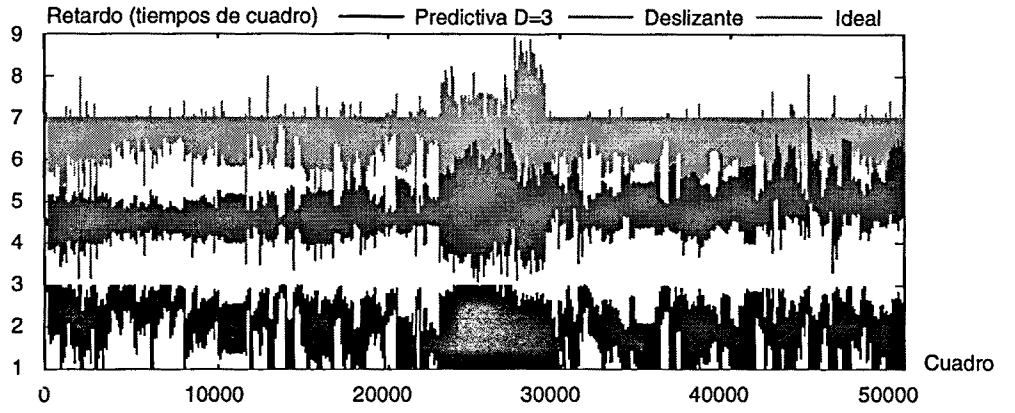

Figura 4.46. Comparación de retardos para la serie completa "Geografía de Cataluña"

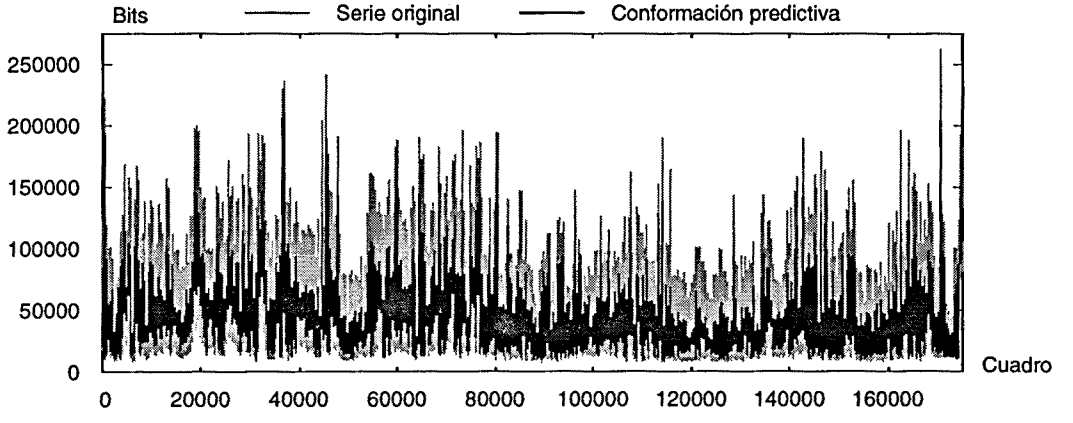

Figura 4.47. Conformación predictiva de la serie completa "Jurassic Park"

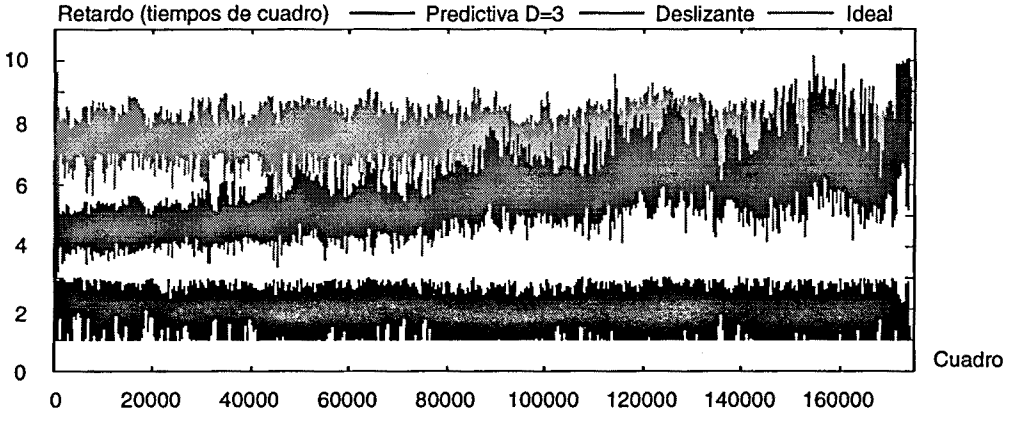

Figura 4.48. Comparación de retardos para la serie completa "Jurassic Park"