Capítulo 4.
Estándares de Interoperabilidad de los sistemas de Videoconferencia

El mercado del facsímil estuvo restringido por muchos años porque las unidades de fax manufacturadas por diferentes vendedores no eran compatibles. Es claro que la explosión del facsímil que ahora experimentamos esta directamente relacionada a el estándar desarrollado por el grupo 3 del Comité Consultivo Internacional para la Telefonía y Telegrafía (CCITT), el cual hace posible que las unidades de fax de diferentes fabricantes sean compatibles. Algo similar ocurrió con la videoconferencia/videoteléfono. El mercado de la videoconferencia punto a punto estuvo restringido por la falta de compatibilidad hasta que surgió la recomendación de CCITT H.261 en 1990, con lo que el mercado de la videoconferencia ha crecido enormemente. Hay otros tres factores que han influido en este crecimiento, el primero es el descubrimiento de la tecnología de videocompresión, a partir de la cual, el estándar esta basado. Mediante la combinación de las técnicas de la codificación predictiva, la transformada discreta del coseno (DCT), la compensación de movimiento y la codificación de longitud variable, el estándar hace posible transmitir imágenes de TV de calidad aceptable con bajos requerimientos de ancho de banda, anchos de banda que se han reducido lo bastante para lograr comunicaciones de bajo costo sobre redes digitales conmutadas [Schaphorst 1996].

El segundo factor que ha influido es el desarrollo de la tecnología VLSI la cual redujo los costos de los codecs de video. Ahora en el mercado se encuentran chips mediante los cuales se pueden implantar las tecnologías DCT y de compensación de movimiento, partes del estándar.

El tercer factor es el desarrollo de ISDN (*Red Digital de Servicios Integrados*), la cual promete proveer de servicios de comunicaciones digitales conmutados de bajo costo. El acceso básico de ISDN consiste de dos canales full dúplex de 64 Kbps denominados canales B y un canal también full dúplex de 16 Kbps denominado D. El estándar H.261 está basado en la estructura básica de 64 Kbps de ISDN. Esta da nombre al título de la recomendación H.261 "Video Codec para servicios audiovisuales a PX64 Kbps", donde P es igual a 1, 2,...,etc. Aunque tomará varios años para que ISDN esté disponible globalmente, los video codecs que cumplen con el estándar H.261 pueden ya operar sobre las redes de comunicaciones actualmente disponibles [Martin 1997].

La CCITT es parte de la Organización de la Naciones Unidas, y su propósito es el desarrollo formal de "recomendaciones" para asegurar que las comunicaciones mundiales sean establecidas efectiva y eficientemente. La CCITT trabaja en ciclos de 4 años, y al final de cada periodo se publica un grupo de recomendaciones. Los libros "rojo" y "azul" que contienen estas recomendaciones fueron publicados en 1984 y 1988 respectivamente. En el libro rojo de 1984 se establecieron las primeras recomendaciones para codecs de videoconferencia (la H.120 y H.130) [Ornelas y Díaz 1998]. Estas recomendaciones se definieron específicamente para la región de Europa (625 líneas; 2.048 Mbps, ancho de banda primario) y para la interconexión entre Europa y otras regiones. Debido a que no existían recomendaciones para las regiones fuera de Europa, la CCITT designó un "grupo de especialistas en Codificación para Telefonía Visual" con el fin de desarrollar una recomendación internacional. La CCITT estableció dos objetivos para el grupo de especialistas: 1) Desarrollar una recomendación para un video codec para aplicaciones de videoconferencia que operará a NX384 Kbps (N=1, 2, hasta 5), y 2) Empezar un proceso de estandarización para el video codec de videoconferencia/videoteléfono que operara a MX64 Kbps (M=1,2). El resultado fue una sola recomendación que se aplica a los rangos desde 64 Kbps hasta 2 Mbps, utilizando PX64 Kbps, donde los valores claves para P son 1, 2, 6, 24 y 30 [Ornelas y Díaz 1998].

En 1989, un diverso número de organizaciones en Europa, EUA y Japón desarrollaron codecs flexibles para encontrar una especificación preliminar de la recomendación. Varios sistemas fueron interconectados en los laboratorios y a través de largas distancias para poder validar la recomendación. Estas pruebas resultaron exitosas y apareció entonces una versión preliminar de la recomendación H.261 en el libro azul de CCITT. Sin embargo, esta versión estaba incompleta, la versión final de la recomendación fue aprobada en diciembre de 1990.

Actualmente, la mayoría de los fabricantes ofrecen algoritmos de compresión que cumplen con los requisitos especificados en la norma CCITT H.261, y ofrecen también en el mismo codec, algoritmos de compresión propios. La norma CCITT H.261 proporciona un mínimo común denominador para asegurar la comunicación entre codecs de diferentes fabricantes [Ornelas y Díaz 1998]. A continuación se listan las recomendaciones de la CCITT que definen a las comunicaciones audio visuales sobre redes digitales de banda angosta.

Servicios

1. F.710 Servicios de Videoconferencia.
2. F.721 Servicio Básico de videoteléfono en banda angosta en la ISDN.
3. H.200 Recomendaciones para servicios audiovisuales.

Equipo Terminal Audio Visual: punto a punto.

PX64

1. H.320 Equipo terminal y sistemas de telefonía visual para banda angosta.
2. H.261 Video codec para servicios audiovisuales a PX64 Kbps.
3. H.221 Estructura de comunicaciones para un canal de 64 Kbps a 1920 Mbps en teleservicios audiovisuales.
4. H.242 Sistemas para el establecimiento de las comunicaciones entre terminales audiovisuales usando canales digitales arriba de 2 Mbps.
5. H.230 Control de sincronización y señales de indicación para sistemas audiovisuales.
   Audio
6. G.711 Modulación por codificación por pulsos (MPC) de frecuencias de voz.
7. G.722 Codificación de audio de 7 Khz dentro de 64 Kbps. Diversos
8. H.100 Sistemas de telefonía visual.
9. H.110 Conexiones hipotéticas de referencia utilizando grupos primarios de transmisiones digitales.
10. H.120 Codecs para videoconferencia para grupos primarios de transmisiones digitales.
11. H.130 Estructuras para la interconexión internacional de codecs digitales para videoconferencia de telefonía visual.

Multipunto

1. H.231 Unidades de control de multipunto (MCU) para sistemas audiovisuales usando canales digitales de mas de 2 Mbps.
2. H.243 Procedimientos básicos para el establecimiento de las comunicaciones entre tres o más terminales audiovisuales usando canales digitales de mas de 2 Mbps.

Seguridad

H.233 Recomendaciones para sistemas de confiabilidad para servicios audiovisuales.

H.KEY Recomendaciones de la CCITT de encriptación para servicios

audiovisuales.

Recomendaciones de la CCITT que definen las comunicaciones audiovisuales sobre ISDN de banda ancha (B-ISDN).

H.26x Video codecs para servicios audiovisuales a velocidades que incluyen a B-ISDN.

Estándares ISO para almacenamiento y utilización de material audiovisual (MPEG).

Codificación de imágenes con movimiento y medios de almacenamiento digital para video para mas de 1.5 Mbps (MPEG1:Comité 11172).

Codificación de imágenes con movimiento y medios de almacenamiento digital para video para mas de 10 Mbps (MPEG2).

Codificación de imágenes con movimiento y medios de almacenamiento digital para video para mas de 40 Mbps (MPEG3).

Estándar ISO para compresión de imágenes fijas (JPEG).

Compresión digital y codificación de imágenes fijas.

Compresión ISO Bi-nivel compresión de imágenes fijas.

Estándar de compresión progresiva bi-nivel para imágenes.

4.1 Tipos de compresión y decodificación

La forma más común de la señal de video todavía es la señal analógica. Esta señal se obtiene a través de un proceso conocido como de búsqueda (scanning). Este proceso graba los valores de intensidad de la señal espacio-temporal en la dirección h, convirtiéndola en una señal unidimensional. Esta señal se señaliza con pulsos de sincronización verticales y horizontales para conseguir la señal de video final.

Existen dos tipos de búsquedas que pueden ser progresivo o entralazado:

La progresiva busca todas las líneas horizontales para formar el cuadro (frame) completo y se usa en la industria de los monitores de computadoras.

La búsqueda entrelazada se utiliza en la industria de TV. Aquí, las líneas horizontales pares e impares de un cuadro se buscan de forma separada consiguiendo los dos campos de un cuadro.

Existen principalmente tres estándares de video analógico, estas son

denominadas:

Video Compuesto.

Video RGB o Componente y,

S-Video ó S-VHS.

En el formato de video compuesto, la componente luminancia y las dos de crominancia son codificadas juntas como una única señal.

En contraposición está el formato RGB o Componente en el que se codifican por separado, y cada componente tiene un canal para ella.

En el formato S-Video, también conocido como Y/C, existen dos señales independientes, una de ellas contiene únicamente la información de luminancia, mientras que el segundo canal contiene la información de crominancia C (U y V). El estándar de video compuesto incluye el formato NTSC utilizado en USA y Japón, y PAL/SECAM utilizado en Europa.

Tabla 4.1 Equivalencia entre señales [Kochervar 1993]

Y =	0.30R	+ 0.59G	+ 0.11B
U =	-0.15R	- 0.29G	+0.47B
V =	0.62R	- 0.52G	- 0.10B

4.2 Relación espacio – temporal: Proceso de búsqueda

4.2.1 Estándares para video digital

La migración hacia la tecnología digital ha estado acompañada por una evolución de los estándares de video digital para varias aplicaciones. CCIR define el estándar para la industria de TV, mientras que VESA los define para la industria de las computadoras.

Tabla 4.2 Estándares TV

PARAMETROS	CCIR 601 NTSC	CCIR 601 PAL
Pixeles/Línea (L)	720	720
Líneas/Imagen (L)	485	576
Frecuencia temporal	60 campos/s	50 campos/s
Relación de aspecto	4:3	4:3
Entrelazado	2:1	2:1

Tabla 4.3 Estándares monitores

PARAMETROS	VGA	TARGA	S
Pixeles/Línea	640	512	8
Líneas/Imagen	480	480	4
Frecuencia temporal	72	72	7
Entrelazado	1:1	1:1	1

4.3 La compresión

Los avances dramáticos registrados en las tecnologías del procesamiento de señales y VLSI registrados en la pasada década han traído progresos significativos en el desarrollo de tecnologías de compresión para señales de video a diferentes velocidades de transmisión. De esta manera, han emergido codificadores de video que en un tiempo eran técnicamente o

económicamente imposibles, y hoy en día son herramientas prácticas [Jeffay, Stone y Smith 1994].

La información de video contiene una serie de imágenes ó "cuadros" y el efecto del movimiento es llevado a cabo a través de cambios pequeños y continuos en los cuadros. Debido a que la velocidad de estas imágenes es de 30 cuadros por segundo, los cambios continuos entre cuadros darán la sensación al ojo humano de movimiento natural [Hopper 1994]. Las imágenes de video están compuestas de información en el dominio del espacio y el tiempo [Draoli et al. 1995]. La información en el dominio del espacio es provista en cada cuadro, y la información en el dominio del tiempo es provista por imágenes que cambian en el tiempo (por ejemplo, las diferencias entre cuadros). Puesto que los cambios entre cuadros colindantes son diminutos, los objetos parecen moverse suavemente.

En los sistemas de video digital, cada cuadro es muestreado en unidades de pixeles ó elementos de imagen. El valor de luminancia (que es la medida de la intensidad de la luz que consiste en el cociente de la intensidad luminosa de una superficie partido por el área aparente de esa superficie) de cada pixel es cuantificado con ocho bits por pixel para el caso de imágenes blanco y negro. En el caso de imágenes de color, cada pixel mantiene la información de color asociada; por lo tanto, los tres elementos de la información de luminancia designados como rojo, verde y azul, son cuantificados a ocho bits. La información de video compuesta de esta manera posee una cantidad tremenda de información; por lo que, para transmisión o almacenamiento, se requiere de la compresión (o codificación) de la imagen [Gaibisso et al. 1994].

La técnica de compresión de video consiste fundamentalmente de tres pasos [Jeffay et al. 1994]. Primero, el preprocesamiento de las diferentes fuentes de video de entrada (señales de TV, señales de televisión de alta definición HDTV, señales de videograbadoras VHS, BETA, S-VHS, etc.), paso en el cual se realiza el filtrado de las señal de entrada para remover componentes no útiles y el ruido que pudiera haber en esta. El segundo paso es la conversión de la señal a un formato intermedio común (CIF), y por último el paso de la compresión. Las imágenes comprimidas son transmitidas a través de la línea de transmisión digital y se hacen llegar al receptor donde son reconvertidas a el formato común CIF y son desplegadas después de haber pasado por la etapa de post-procesamiento.

Mediante la compresión de la imagen se elimina información redundante, principalmente la información redundante en el dominio de espacio y del tiempo. En general, las redundancias en el dominio del espacio son debidas a las pequeñas diferencias entre pixeles contiguos de un cuadro dado, y aquellas dadas en el dominio del tiempo son debidas a los pequeños cambios dados en cuadros contiguos causados por el movimiento de un objeto. El método para eliminar las redundancias en el

dominio del espacio es llamado codificación de cuadros, la cual puede ser dividida en codificación por predicción, codificación de la transformada y codificación de la sub-banda [Schulzrinne 1994]. En el otro extremo, las redundancias en el dominio del tiempo pueden ser eliminadas mediante el método de codificación de intercuadros, que también incluye los métodos de compensación/estimación del movimiento, el cual compensa el movimiento a través de la estimación del mismo.

4.3.1 La solución: Compresión

La solución lógica a este problema es la compresión digital. Compresión implica disminuir el numero de parámetros requerido para representar la señal, manteniendo una buena calidad visual. Estos parámetros son codificados para almacenarse o transmitirse. El resultado de la compresión de video digital es que se convierte a un formato de datos que puede transmitirse a través de las redes de comunicaciones actuales y ser procesadas por computadoras.

Para entender el proceso de compresión es importante reconocer las diferentes redundancias presentes en los parámetros de una señal de video:

Espacial

Temporal

Psicovisual

Codificación

La redundancia espacial ocurre porque en un cuadro individual los pixeles cercanos (contiguos) tienen un grado de correlación, es decir, son muy parecidos (por ejemplo, en una imagen que muestre un prado verde bajo un cielo azul, los valores de los pixeles del prado serán muy parecidos entre ellos y del mismo modo los del cielo). Los pixeles en cuadros consecutivos de una señal también están correlacionados, determinando una redundancia temporal (si la señal de video fuera un recorrido por el prado, entre una imagen y la siguiente abría un gran parecido). Además, el sistema de visión humano no trata toda la información visual con igual sensibilidad, lo que determina una redundancia psicovisual (por ejemplo, el ojo es más sensible a cambios en la luminancia que en la crominancia).

El ojo es también menos sensible a las altas frecuencias. Por lo tanto, un criterio que toma mucha importancia es estudiar como percibe el ojo humano la intensidad de los pixeles para así, dar mayor importancia a unos u otros parámetros.

Finalmente, no todos los parámetros ocurren con la misma probabilidad

en una imagen. Por lo tanto resulta que no todos necesitarán el mismo número de bits para codificarlos, aprovechando la redundancia en la codificación.

Durante los últimos años han emergido diferentes estándares de compresión, incluyendo algunos propietarios, dirigidos a diversas aplicaciones con diferentes necesidades de velocidad. Por ejemplo, la recomendación ITU H.261, también conocida como el estándar px64 ha surgido para aplicaciones de video conferencia. MPEG, Moving Picture Expert Group, es un comité del organismo ISO e IEC que es responsable de los estándares MPEG-1, MPEG-2, y los actuales MPEG-4 y MPEG-7 aún en fase de especificación.

Los estándares MPEG son genéricos y universales en el sentido que simplemente especifican una sintaxis de la trama para el transporte de los datos obtenidos mediante los algoritmos de compresión de video y audio, no estando definidos los procesos de compresión (lo que permite plena libertad en su realización).

4.3.2 Compresión MPEG

En la especificación MPEG-1 y MPEG-2 existen tres partes diferenciadas, llamadas, Sistema, Video y Audio. La parte de video define la sintaxis y la semántica del flujo de bits de la señal de video comprimida. La parte de audio opera igual, mientras que la parte Sistema se dirige al problema de la multiplexación de audio y video en un único flujo de datos con toda la información necesaria de sincronismo, sin desbordar los buffers del decodificador.

Adicionalmente, MPEG-2 contiene una cuarta parte llamada DSMCC (Digital Storage Media Command Control), que define un conjunto de protocolos para la recuperación y almacenamiento de los datos MPEG desde y hacia un medio de almacenamiento digital.

4.3.3Video

Se examinará ahora la estructura de un flujo de video no escalable para entender la compresión de video. En el proceso, se verá como el algoritmo es realmente un conjunto de herramientas que individualmente explotan las diferentes redundancias. En el nivel más alto, el flujo de datos de video consiste en secuencias de video. MPEG-1 sólo soporta secuencias progresivas, mientras que MPEG-2 permite secuencias progresivas y entrelazadas. Cada secuencia de video consiste en un numero variable de grupos de imágenes (GOP, group of pictures). Un GOP contiene un número variable de imágenes y jugará un papel muy importante en el proceso de compresión.

4.3.4 Grupo de imágenes (GOP)

Una imagen puede ser un cuadro o un campo de una imagen (sólo en MPEG-2). A partir de este momento se hablará de imagen, frame, cuadro o campo de forma indistinta.

Matemáticamente, cada imagen es realmente una unión de los valores que representan a un pixel: una componente de luminancia y dos de crominancia; es decir, tres matrices de pixeles. Ya que el ojo humano no es muy sensible a los cambios de la región cromática comparada con la región de luminancia, las matrices de croma son decimadas o reducidas en tamaño por un factor de dos en ambas direcciones horizontal y vertical.

Consecuentemente hay una cuarta parte de números de pixeles de crominancia para procesar con los pixeles de luminancia. Este formato, denominado formato (4:2:0), se emplea en MPEG-1.

MPEG-2 adicionalmente permite la posibilidad de no decimar o sólo decimar horizontalmente la componente croma, consiguiente formatos 4:4:4 y 4:2:2 respectivamente.

Las imágenes pueden clasificarse principalmente en tres tipos basados en sus esquemas de compresión.

I (Intraframes) o intra cuadros.

P(Predictive) o cuadros predecidos.

B(Bi-directional) o cuadros bidireccionales.

Las imágenes **I** son codificadas por ellas mismas, de ahí el nombre intra. La técnica de codificación para estas imágenes entra en la categoría de la codificación por transformada. Cada imagen se divide en bloques de pixeles de 8x8 no solapados. Cuatro de estos bloques se organizan adicionalmente en un bloque mayor de tamaño 16x16, llamado macrobloque.

La Transformada Discreta Coseno se aplica a cada bloque de 8x8 individualmente. La transformada explota la correlación espacial de los pixeles convirtiéndolos en un conjunto de coeficientes independientes. Los coeficientes de baja frecuencia contienen más energía que los de alta frecuencia. Estos coeficientes son cuantificados utilizando una matriz de cuantificación, este proceso permite que los coeficientes de baja frecuencia (contienen gran energía) sean codificados con un número mayor de bits, mientras que para los coeficientes de mayor frecuencia (menor energía) se usan menos bits o cero bits.

Los coeficientes de alta energía pueden eliminarse ya que el ojo carece de la habilidad de detectar cambios de alta frecuencia. Reteniendo sólo un subconjunto de los coeficientes se reduce el número total de parámetros necesarios para la representación en una cantidad considerable. El proceso es idéntico para los bloques de pixeles de luminancia y crominancia. Sin embargo, ya que la sensibilidad del ojo humano a la luminancia y a la crominancia varía, las matrices de cuantificación para las dos varían.

El proceso de cuantificación también ayuda en el control de velocidad, por ej. permitiendo al codificador producir un flujo de bits a una determinada velocidad. Los coeficientes DCT son codificados empleando una combinación de dos esquemas de codificación especiales: Run length y Huffman. Los coeficientes son escaneados siguiendo un patrón en *zig-zag* para crear una secuencia de una dimensión. MPEG-2 proporciona un patrón alternativo. La secuencia resultante de 1-D usualmente contiene un gran número de ceros debido a la naturaleza del espectro DCT y del proceso de cuantificación. Cada coeficiente diferente de cero se asocia con un par de apuntadores. Primero, su posición en el bloque que se indica por el número de ceros entre él y el coeficiente anterior diferente de cero (run length). Segundo, su valor.

4.3.5 Zig-Zag

Basado en estos dos apuntadores, se le asigna un código de longitud variable (Huffman) en función de una tabla predeterminada. Este proceso se realiza de tal forma que las combinaciones con una alta probabilidad consiguen un código con pocos bits, mientras que los poco habituales obtienen un código mayor. Adoptando esta codificación sin pérdidas, el número total de bits disminuye. Sin embargo, ya que la redundancia espacial es limitada, las imágenes **I** sólo proporcionan una compresión moderada. Estas imágenes son muy importantes para acceso aleatorio utilizado para fines de edición. La frecuencia de imágenes **I** es normalmente de una cada 12 o 15 cuadros o frames. Un GOP está delimitado por dos cuadros **I** .

En las imágenes P y B es donde MPEG proporciona su máxima eficiencia en compresión. Esto lo consigue mediante una técnica llamada predicción basada en la compensación de movimiento (MC: Motion Compensation), que explota la redundancia temporal. Ya que los cuadros están relacionados, podemos asumir que una imagen puede ser modelada como una translación de la imagen en el instante anterior. Entonces, es posible representar de manera precisa o predecir los valores de un cuadro basándonos en los valores del cuadro anterior, estimando el movimiento. Este proceso disminuye considerablemente la cantidad de información. En las imágenes P, cada macrobloque de tamaño 16x16 se predice a partir de un macrobloque de la anterior imagen I. Ya que, los cuadros son

instantáneos en el tiempo de un objeto en movimiento, los macrobloques en los dos cuadros pueden no corresponder a la misma localización espacial, por lo tanto, se debe proceder a buscar en el cuadro I para encontrar un macrobloque que coindica lo máximo posible con el macrobloque que se está considerando en el cuadro P.

La diferencia entre los dos macrobloques es el error de predicción. Este error puede codificarse como tal o en el dominio DCT. La DCT del error consigue pocos coeficientes de alta frecuencia, que tras la cuantificación requieren un numero menor de bits para su representación. Las matrices de cuantificación para los bloques de error de predicción son diferentes de las utilizadas en los intra bloques, debido a la distinta naturaleza de sus espectros. La distancia en las direcciones horizontal y vertical del macrobloque coincidente con el macrobloque estimado se denomina vector de movimiento.

4.4 Proceso de predicción por compensación de movimiento

Los vectores de movimiento representan la translación de las imágenes de los bloques entre cuadros. Estos vectores se necesitan para la reconstrucción y son codificados de forma diferencial en el flujo de datos. Se utiliza codificación diferencial ya que reduce el total de bits requeridos para transmitir la diferencia entre los vectores de movimiento de los cuadros consecutivos. La eficiencia de la compresión y la calidad de la reconstrucción de la señal de video depende de la exactitud en la estimación del movimiento.

El método para este cálculo no se especifica en el estándar y por lo tanto está abierto a diferentes implementaciones y diseños, aunque evidentemente existe una relación directa entre la exactitud de la estimación de movimiento y la complejidad de su cálculo.

Para los cuadros B, se utiliza la predicción de la compensación de movimiento y la interpolación usando cuadros de referencia presentes antes o después de ellos, donde las referencias pueden ser cuadros I y P.

La predicción no es casual, ya que se usan cuadros anteriores y posteriores. Comparados con los cuadros I y P, los B proporcionan la máxima compresión. Otras ventajas de los cuadros B son la reducción del ruido debido a un proceso de promedio y el uso de cuadros posteriores para la codificación. Esto es particularmente útil para la codificación de ”áreas descubiertas”. Los cuadros B nunca se usan por sí solos para predicciones para no propagar errores. MPEG-2 permite MC para cuadros y campos. Para una secuencia de imágenes de variación lenta es mejor codificar los cuadros (combinando los dos campos, si es necesario). MC

basada en campos es especialmente útil cuando la señal de video incluye movimientos rápidos.

4.5 Herramientas de compresión

Como se ha visto los algoritmos de compresión MPEG son una combinación inteligente de un número de diversas herramientas, cada una de ellas explota una redundancia concreta de la señal de video (tabla 4.4).

Tabla 4.4 Herramientas de compresión

HERRAMIENTA	REDUNDANCIA
DCT	Espacial
Predicción de compensación de movimiento	Temporal
Codificación Run Length/Huffman	Codificación
Codificación diferencial	Temporal

4.6 MPEG-2

A la hora de almacenar video, un método que incrementa de forma significativa la eficiencia de la compresión MPEG es la utilización de una velocidad variable de bits (VBR, Variable Bit Rate). Este método ofrece la posibilidad de adaptar la velocidad utilizada por el codificador a la complejidad de la imagen en segmentos de 25ms. Por ejemplo, imágenes simples necesitaran una velocidad instantánea de bits baja, mientras que una compleja demandará una velocidad mayor. Por el contrario si utilizamos una velocidad constante (FBR, Fixed Bit Rate), esta será aquella necesaria para codificar la imagen más compleja y por lo tanto en el resto de casos se desperdiciará espacio. La codificación a velocidad constante es inherentemente un subconjunto de la codificación VBR, por lo que todos los decodificadores soportarán FBR, siendo VBR opcional.

Tabla 4.5 Definición de herramientas según el tamaño de la imagen, velocidad de cuadros o de transmisión

NIVEL	PARÁMETROS			
	Muestras/línea	líneas/cuadro	Cuadros/s	Max. vel. Mbps
HIGH	1920	1152	60	80
HIGH 1440	1440	1152	60	60
MAIN	720	576	30	15
LOW	352	288	30	4

Tabla 4.6 Perfil principal para la definición de la compresión

PERFIL	CARACTERÍSTICAS
MAIN	Soporta algoritmos de codificación no escalables para vide progresivo/entrelazado Soporta predicción de cuadros B A aleatorio Representación 4:2:0 YUV (4:1:1)
SNR Escalable	Soporta toda la funcionalidad de MAIN Codificación escala SNR
Espacial Escalable	Soporta toda la funcionalidad de SNR Escalable Codificació espacial escalable Representación 4:0:0
HIGH	Soporta toda la funcionalidad del perfil Espacial escalable capas con modos de codificación escalable SNR y Espacial Representación 4:2:2

SIMPLE	Soporta toda la funcionalidad de MAIN excepto la predicci cuadros B

4.7 Codificación de audio digital (MPEG y Dolby AC-3)

Los métodos de codificación de audio que existen en la actualidad se basan en algoritmos de compresión y en codificación multicanal.

Los algoritmos de compresión de audio se fundamentan en aspectos perceptibles al oído humano. Básicamente son dos los fenómenos que son objeto de estudio y que han originado los métodos de compresión:

La curva de sensibilidad del oído

El fenómeno de enmascaramiento

El oído humano detecta sonidos entre 20Hz y 20KHz. Pero su sensibilidad depende de la frecuencia del sonido, de esta forma, dos frecuencias con la misma potencia son interpretadas por nuestro oído de forma diferente, teniendo la sensación de que una es más fuerte que otra, o incluso, oír una y no la otra. La curva que indica cual es la potencia mínima (umbral) que nuestro oído detecta es la curva de sensibilidad:

MPEG define 3 capas de codificación de audio, cada una añade complejidad a la anterior. La codificación se realiza dividiendo las secuencias de audio en tramas (de 384 muestras), que se filtra para obtener las bandas críticas:

La capa 1 sólo considera en enmascaramiento frecuencial,

La capa 2 considera además el enmascaramiento temporal estudiando 3 tramas a la vez,

La capa 3 utiliza filtros no lineales, elimina redundancias provocadas por el muestreo y utiliza codificación de Huffman.

Este esquema de codificación es prácticamente idéntico para MPEG-2 y para DOLBY AC-3, aunque cada una de ellas sugiere que es mejor que la otra. DOLBY AC-3 se basa en ser la primera que desarrolló un sistema multicanal de audio, mientras que MPEG en su poder de integración de estándares. El resultado es la incompatibilidad explícita entre ambos formatos, aunque el mercado multimedia se han adaptado rápidamente soportando ambos formatos (como por ejemplo, DVD). La gran diferencia

entre estos sistemas es la ecualización de cada canal para crear diferentes perfiles de ”envolvencia”.

4.8 Curva de sensibilidad (típica del oído)

Podemos observar que nuestro oído es muy sensible a frecuencias entre 2 y 4KHz (aproximadamente). Además observamos que si la potencia de una cierta frecuencia no supera el umbral de la sensibilidad del oído, simplemente no la oiremos, por lo tanto no hace falta que la codifiquemos. Este es un primer paso en la compresión: eliminar las señales que no oiremos.

Existe otro tipo de señales que tampoco oímos: aquellas que son enmascaradas. Imaginemos una señal de 1KHz con un potencia tal que supera el umbral y que, por lo tanto, oímos. Si aparece de forma simultanea otra señal de 0.5KHz y vamos aumentando su potencia llegará un instante en el que no oiremos la señal de 1KHz ya que ha sido enmascarada. Esto se debe a que la potencia de una señal hace que la sensibilidad del oído varíe, necesitando más potencia de las señales próximas en frecuencia para poder oírlas.

4.9 Fenómeno de enmascaramiento

El enmascaramiento gana importancia cuando los sonidos son cercanos en frecuencia y la frecuencia enmascaradora es inferior que la enmascarada. Para poder cuantificar el fenómeno de enmascaramiento surge el concepto de banda crítica como el ancho de banda máxima alrededor de una frecuencia para que no haya enmascaramiento, por lo tanto, sólo se produce éste entre bandas contiguas. Además, estas bandas están distribuidas siguiendo una escala logarítmica, simulando la escala perceptiva del oído. Una escala de medida perceptual es la escala BARK que relaciona las frecuencias acústicas con la resolución perceptual de éstas [Coelho et al. 1997].

4.9.1 Escala perceptual BARK

A partir de esta escala de bandas de frecuencia y de un modelo psicoacústico se determinará que frecuencias se enmascaran y cuales no.

Además existe enmascaramiento temporal: cuando oímos un sonido de alta potencia y para de pronto, seguimos oyéndolo durante un breve instante de tiempo que puede enmascarar a otras señales.

El proceso de compresión es el siguiente:

1. Se divide la señal de audio en bandas frecuenciales mediante filtros convolucionales de tal forma que se corresponden con 32 bandas críticas (aproximadamente). Filtrado subbanda.

2. Se determina el umbral de potencia de cada banda crítica considerando el fenómeno de enmascaramiento por las bandas contiguas a partir de un modelo psicoacústico.

3. Si la potencia de una banda es menor que el umbral no se codifica.

4. En caso contrario, se determina el número de bits necesario para representar el coeficiente tal que el ruido introducido en la cuantificación sea menor que el efecto de enmascaramiento.

5. Se crea la trama de datos:

4.9.2 Esquema de codificación

Por ejemplo, imaginemos que tras el análisis correspondiente se encuentra que los niveles de potencia de las bandas son:

Banda 1 2 3 4 5 6 7 8 9 10 ...

Nivel (db) 0 8 12 10 6 2 10 60 35 20 ...

Si el nivel en la octava banda es de 60dB y según el modelo psicoacústico provoca un enmascaramiento de 12dB sobre la banda 7 y 15dB sobre la banda 9,

El nivel en la banda 7 es 10dB (<12dB), por lo tanto la enmascara y se ignora.

El nivel en la banda 9 es de 35dB (> 15dB), por lo tanto se codifica.

4.9.3 Sistema multicanal

Todo el mundo esta familiarizado con el sonido estéreo, en el que el sonido se reproduce usando dos canales, derecho e izquierdo. Pronto se descubrió que para algunas señales, como por ejemplo diálogos en películas, la adición de un tercer canal central proporcionaba una mejor localización dentro de la escena. Para dar una sensación espacial como en una sala de cine, se pensó en un cuarto canal "sorround", de ancho de banda limitado, que se reproducía en dos altavoces localizados detrás del

público [Heman 1997].

El siguiente paso era proporcionar la sensación de sonido envolvente, esto se consiguió con dos canales sorround separados.

Además se pensó en un sexto canal para proporcionar un canal de baja frecuencia, a este sexto canal se le denominó LFE (Low Frequency Enhancement). Al sistema resultante se le denominó 5+1. El subwoofer usado para reproducir el canal LFE no necesita una localización particular, ya que nuestro oído tiene una capacidad muy limitada para detectar la dirección de las frecuencias bajas (20-120Hz).

4.9.4 Sistema con 5 canales

MPEG-2 y DOLBY AC-3 proporcionan este sistema de 5+1. DOLBY utiliza una ecualización propietaria resultando el sistema DOLBY SORROUND PRO LOGIC. El Dolby Surround Pro Logic, o sonido envolvente, tiene gran difusión en equipos reproductores de música, necesitando un amplificador Dolby Surround, cinco altavoces y una fuente estéreo. El sistema MPEG-2 actualmente "sólo" se utiliza para la difusión vía satélite, cable y para el formato DVD (junto con DOLBY AC-3).

El sistema MPEG-2 proporciona dos canales más pensando en locales de grandes dimensiones (cines,...) para cubrir ángulos muertos. El sistema MPEG-2 está basado en la compatibilidad, ya que permite la reproducción en sistemas que sólo soporten un número de canales limitado. Esta compatibilidad se consigue empleando técnicas de multiplexación matricial durante la codificación y decodificación.

Por lo que se ha visto a lo largo de esta capítulo se puede concluir que la compresión es una excelente herramienta para la transmisión de video, dentro de los ambientes distribuídos, heterogéneos y multimedios que engloban a la videoconferencia.

En el siguiente capítulo se propone un modelo conceptual contrastándolo con dos modelos de comunicación visual existentes. Este es un modelo real que engloba a las tres áreas mencionadas en los capítulos 2,3 y 4 – bibliotecas digitales, videoconferencia y los estándares de interoperabilidad de los sistemas de videoconferencia.

Morales Salcedo, R. 1999. **Aplicaciones de la Videoconferencia en Bibliotecas Digitales**. Tesis Maestría. Ciencias con Especialidad en Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas Puebla. Mayo. Derechos Reservados © 1999.