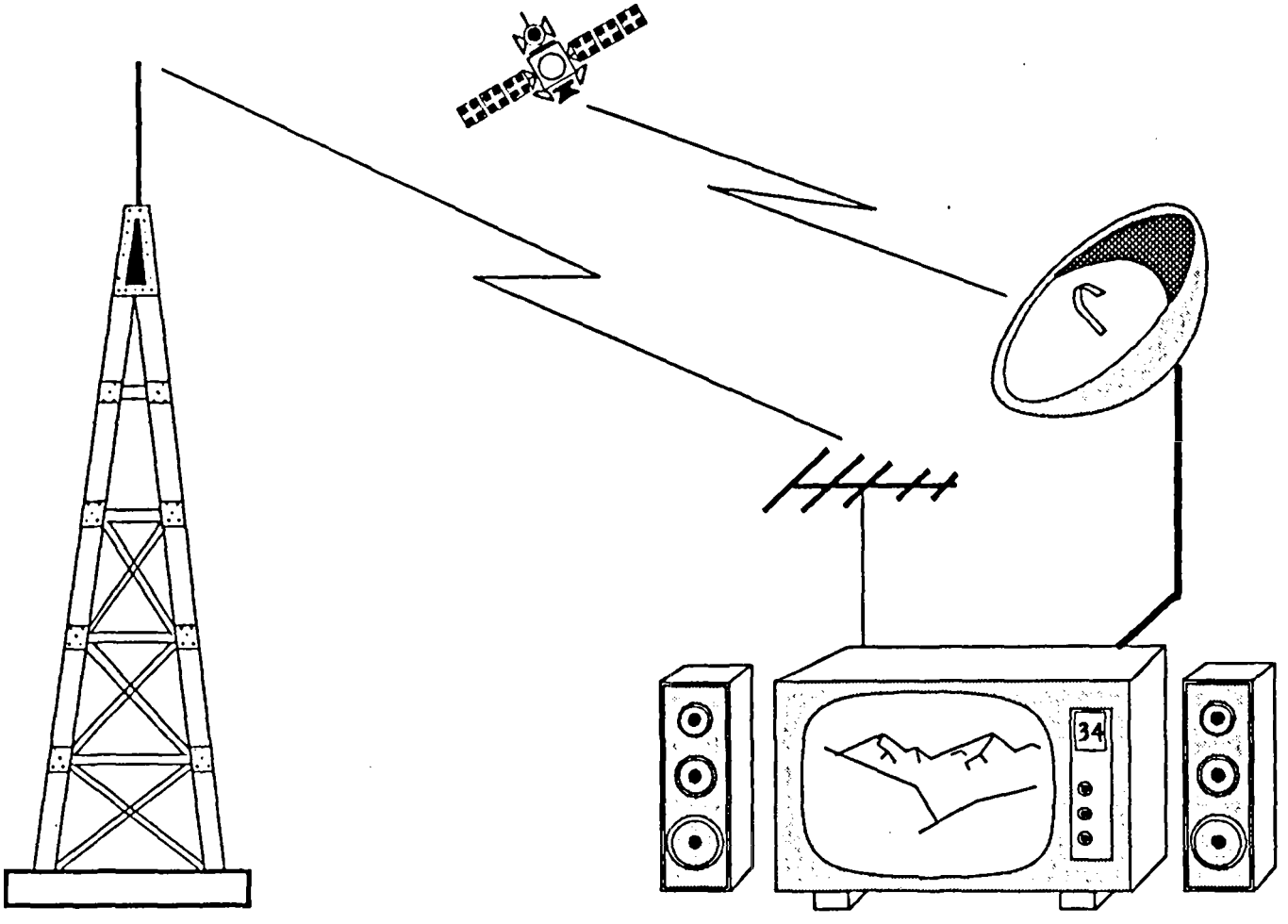
UNIÓN INTERNACIONAL DE TELECOMUNICACIONES

1992 - RECOMENDACIONES DEL CCIR

(Nuevas y revisadas con fecha 15 de septiembre de 1992)



Serie RBT

SERVICIO DE RADIODIFUSIÓN (TELEVISIÓN)

CCIR COMITÉ CONSULTIVO INTERNACIONAL DE RADIOCOMUNICACIONES

ISBN 92-61-04593-6

Ginebra, 1992

Recomendacion 500-5 (1992)

Método de Evaluación subjectiva de la Calidad de las Imágenes de Televisión

Un extracto de la publicación:
Recomendaciones CCIR: Serie RBT: Servicio de Radiodifusión (Televisión)
(Ginebra: UIT, 1992), pp. 166-189

RECOMENDACIÓN 500-5

MÉTODO DE EVALUACIÓN SUBJETIVA DE LA CALIDAD DE LAS IMÁGENES DE TELEVISIÓN

(Cuestión 119/11)

(1974-1978-1982-1986-1990-1992)

El CCIR,

considerando

a) que se poseen numerosos datos acerca de los métodos empleados en diversos laboratorios para evaluar la calidad de las imágenes;

b) que el análisis de estos métodos demuestra que existe una gran concordancia entre los diferentes laboratorios acerca de diversos aspectos de estas pruebas;

c) que la adopción de métodos normalizados reviste importancia para el intercambio de información entre laboratorios;

d) que en las evaluaciones, rutinarias o no, de la calidad y/o degradación de la imagen, realizadas por ciertos técnicos supervisores durante las tareas especiales o de rutina, utilizando escalas de cinco notas, pueden utilizarse también ciertos aspectos de los métodos recomendados para la evaluación en laboratorio;

e) que la introducción de nuevos métodos de procesamiento de señales de televisión (como la codificación digital y la reducción de la velocidad binaria), nuevos tipos de señales de televisión que utilizan componentes multiplexados en el tiempo y, posiblemente, nuevos servicios (como la televisión de calidad mejorada y la TVAD) podrían requerir cambios de los métodos de evaluación subjetiva,

recomienda

1. que los métodos generales de prueba, las escalas de apreciación y las condiciones de observación para la evaluación de la calidad de las imágenes descritas a continuación se utilicen para las experiencias de laboratorio y, siempre que sea posible, para las evaluaciones prácticas;

2. que, en un futuro próximo y a pesar de la existencia de otros métodos y del desarrollo de nuevos métodos, deberían utilizarse, cuando fuera posible, los que se describen en los § 2 y 3 del anexo 1 a esta Recomendación;

3. que, dada la importancia que tiene establecer la base de las evaluaciones subjetivas, todos los informes de pruebas deberían suministrar las descripciones más completas posibles de las configuraciones y materiales de prueba, de los observadores y de los métodos.

Nota 1 – En el anexo 1 figura información relativa a los métodos de evaluación subjetiva para determinar la calidad de funcionamiento de los sistemas de televisión.

Nota 2 – El anexo 2 contiene información sobre los métodos de evaluación subjetiva para establecer las degradaciones debidas a la codificación digital de las señales de televisión.

ANEXO 1

1. Introducción

Se utilizan métodos de evaluación subjetiva para determinar la calidad de funcionamiento de sistemas de televisión a través de mediciones que anticipan de manera más directa las reacciones de quienes podrían ver los sistemas probados. En este aspecto, se comprende que no sería posible caracterizar totalmente la calidad de funcionamiento del sistema por medios objetivos; en consecuencia, es necesario complementar las mediciones objetivas con mediciones subjetivas.

En general, existen dos clases de evaluaciones subjetivas. En primer lugar, hay evaluaciones que determinan la calidad de funcionamiento de sistemas bajo condiciones óptimas, los que típicamente se denominan evaluaciones de calidad. En segundo lugar, hay evaluaciones que determinan la capacidad de los sistemas de mantener la calidad en condiciones no óptimas que se relacionan con la transmisión o emisión. Estas se denominan típicamente evaluaciones de degradación.

Para efectuar evaluaciones subjetivas adecuadas, es necesario primero seleccionar, de las diferentes opciones disponibles, aquélla que mejor se adapte a los objetivos y circunstancias del problema de evaluación inmediato. En la práctica, esto requiere decisiones que conducen a la selección de los métodos de prueba, materiales de prueba y condiciones de observación.

1.1 Selección del método de prueba

En la evaluación de las imágenes de televisión se ha utilizado una amplia variedad de métodos de prueba básicos. Sin embargo, en la práctica se deben emplear métodos específicos para abordar determinados problemas de evaluación. En el cuadro 1 se describen los problemas de evaluación característicos y los métodos utilizados para abordar dichos problemas.

CUADRO 1

Selección del método de prueba

Problema de evaluación	Método utilizado	Origen (Recomendación 500)
Medir la calidad de los sistemas con respecto a una referencia	Método de doble estímulo con escala de calidad continua	§ 3
Cuantificar la calidad de los sistemas (cuando no se dispone de referencias)	Método de valoración cuantitativa (1)	§ 4
Comparar la calidad de sistemas alternativos (cuando no se dispone de referencias)	Método de comparación directa *o* método de valoración cuantitativa (1)	§ 4
Identificar factores en los que se observa que los sistemas difieren	Método con escala multidimensional *o* método por análisis de factores	§ 4
Medir diferencias entre sistemas sobre factores específicos	Método multivalente	
Medir la robustez de los sistemas (es decir, características de fallo)	Método de degradación con doble estímulo	§ 2
Cuantificar la robustez de los sistemas (es decir, características de fallo)	Método de valoración cuantitativa (1)	§ 4
Establecer el punto en el cual una degradación se hace visible	Estimación del umbral por el método de elección forzada *o* método de ajuste	§ 4
Determinar si se perciben diferencias en los sistemas	Método de elección forzada	§ 4

(1) Algunos estudios señalan que este método es más estable cuando se dispone de una gama de calidad completa.

1.2 Selección del material de prueba

Se han tomado una serie de planteamientos para establecer las clases de material de prueba requeridos en las evaluaciones de imágenes de televisión. Sin embargo, en la práctica se deben emplear determinadas clases de materiales de prueba para abordar problemas de evaluación específicos. En el cuadro 2 se describen los problemas de evaluación y de materiales de prueba típicos utilizados para abordar esos problemas.

1.3 Selección de las condiciones de observación

En las evaluaciones de imágenes de televisión convencional se ha de utilizar una serie determinada de condiciones de observación. No obstante, se pueden emplear diferentes distancias de observación para evaluaciones normales y críticas (véase el cuadro 3).

CUADRO 2

Selección del material de prueba*

Problema de evaluación	Material utilizado	Origen (Recomendación 500)
Calidad de funcionamiento global con material de uso habitual	General, «crítico pero no en exceso»	§ 2
Capacidad, aplicaciones críticas (por ejemplo, contribución, postprocesamiento, etc.)	Diverso, incluido el material muy crítico para la aplicación probada	Anexo 1
Calidad de funcionamiento de sistemas «adaptables»	Material muy crítico para el esquema «adaptable» utilizado	Anexo 1
Identificar puntos débiles y posibles mejoras	Crítico, material con propiedades específicas	
Identificar factores en los que se aprecia variación en los sistemas	Amplia gama de material muy abundante	
Conversión entre diferentes normas	Crítico por diferencias (por ejemplo, frecuencia de trama)	

* Se sobreentiende que todos los materiales de prueba deberían poder formar parte de los programas de televisión. En los apéndices 1 y 2 de este anexo se pueden obtener mayores directrices para la selección de materiales de prueba.

CUADRO 3

Selección de las condiciones de observación

Problema de evaluación	Condiciones de observación	Origen (Recomendación 500)
Evaluación de sistemas convencionales	Observación en 6 alturas de imagen	§ 2
Evaluación de sistemas convencionales en condiciones críticas	Observación en 4 alturas de imagen para sistemas de 625 líneas y en 4 ó 5 alturas de imagen para sistemas de 525 líneas	§ 2

2. El método de escala de degradación con doble estímulo (Método «UER»)

2.1 Descripción general

Una apreciación típica puede ser aplicable a la evaluación de un nuevo sistema, o del efecto de la degradación debida al trayecto de transmisión. El organizador de la prueba debería empezar por seleccionar material de prueba suficiente para poder hacer una evaluación significativa y determinar las condiciones de prueba. Si se trata de determinar el efecto de la variación de los parámetros, debe elegirse un conjunto de valores de parámetros que abarque la gama de notas de degradación en un pequeño número de etapas prácticamente iguales. Si se evalúa un nuevo sistema, para el que los valores de los parámetros no pueden variar de esa manera, debe añadirse entonces degradaciones adicionales, pero subjetivamente similares, o utilizarse otro método (como el del § 3).

El método con doble estímulo (método UER) es cíclico en la medida en que se muestra al evaluador una imagen de referencia no degradada, y después la misma imagen degradada. A continuación, se le pide que opine sobre la segunda, con la primera en mente. En *sesiones*, que duran hasta media hora, se muestra al evaluador una serie de imágenes o secuencias en orden aleatorio y con degradaciones aleatorias que abarcan todas las combinaciones requeridas. La imagen no degradada se incluye en las imágenes o secuencias que deben evaluarse. Al final de la serie de sesiones, se calcula la nota media para cada condición de prueba y para cada imagen de prueba.

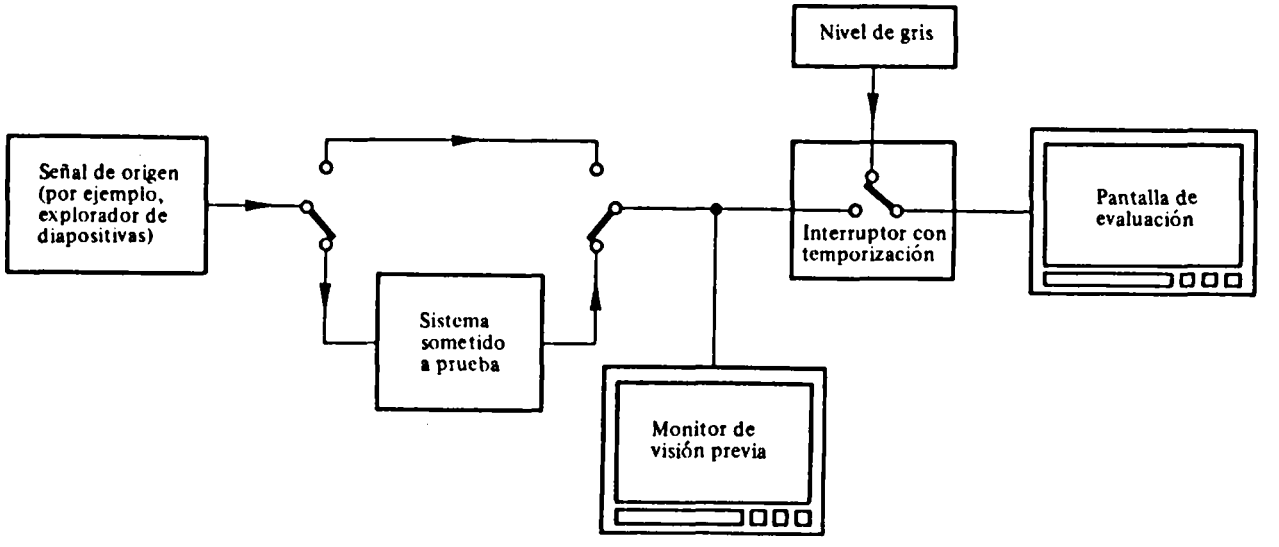
Este método utiliza la escala de degradación, cuyos resultados se suelen considerar más estables para degradaciones pequeñas que para degradaciones considerables. Si bien algunas veces se ha utilizado el método con una escala de degradaciones limitada, es más conveniente utilizarlo con una gama completa de degradaciones.

2.2 *Disposición general*

La disposición general del sistema de prueba debería ser la que se indica en la fig. 1 siguiente.

FIGURA 1

Disposición general de los sistemas de prueba para el método de escala de degradación con doble estímulo



Los evaluadores examinan una imagen de evaluación suministrada por una señal a través de un interruptor con temporización. El trayecto de la señal hacia el interruptor con temporización puede llegar directamente de la señal de origen, o indirectamente a través del sistema sometido a prueba. Los evaluadores examinan una serie de imágenes o de secuencias de prueba. Están dispuestas por pares, de forma que la primera imagen procede directamente del origen, y la segunda es la misma imagen encaminada por el sistema sometido a prueba.

2.3 *Señales de origen*

La señal de origen proporciona directamente la imagen de referencia, y la entrada para el sistema sometido a prueba. Deberá ser de calidad óptima para la norma de televisión utilizada. La ausencia de defectos en la parte de referencia del par presentado es esencial para obtener resultados estables.

Las imágenes y secuencias almacenadas digitalmente son las señales de origen más reproducibles, y son por consiguiente las preferidas. Pueden intercambiarse entre laboratorios, para dar mayor significado a las comparaciones de sistemas. El formato de cinta *D*-1 4:2:2 (Recomendación 657) debería ofrecer una base para el intercambio de imágenes y secuencias de origen cuando se pueda disponer fácil y económicamente de esas máquinas. También se pueden utilizar formatos de cinta de computador.

A corto plazo, los analizadores de diapositivas de 35 mm son la fuente preferida de imágenes fijas, ya que su resolución es adecuada para la evaluación de televisión convencional. La colorimetría y las demás características de las películas pueden dar una apariencia subjetiva distinta de las imágenes de cámara de estudio. Si esto afecta a los resultados, deben utilizarse también fuentes de estudio directas, aunque a menudo sean mucho menos conveniente. Por regla general, los analizadores de diapositivas deberían ajustarse, imagen por imagen, para obtener la mejor calidad subjetiva posible de imagen, ya que esa situación es la que se daría en la práctica.

Las evaluaciones de la capacidad de procesamiento hacia el lado emisión se hacen a menudo con incrustación cromática. En las filmaciones en estudio, la incrustación cromática es muy sensible a la iluminación. Las evaluaciones deberían, pues, usar preferiblemente un par de diapositivas de incrustación cromática especiales, que dieran siempre resultados de alta calidad. En caso necesario, puede introducirse movimiento en la diapositiva de primer plano.

2.4 Condiciones de observación

Las condiciones de observación de los evaluadores deben organizarse como sigue:

2.4.1 Condiciones generales

a)	Relación «distancia de observación/altura de la imagen»	$4H$ y $6H$ *
b)	Valor de cresta de luminancia	70 cd/m^2
c)	Relación entre la luminancia de pantalla inactiva del tubo y el valor de cresta de luminancia	$\leq 0{,}02$
d)	Relación ente la luminancia de la pantalla (cuando sólo se muestra el nivel del negro en una sala completamente a oscuras) y la correspondiente al blanco más intenso	$\approx 0{,}01$
e)	Relación entre la luminancia de fondo detrás del receptor de imágenes y el valor de cresta de luminancia de la imagen	$\approx 0{,}15$
f)	Otra iluminación de la sala	débil
g)	Cromaticidad del fondo	D_{65}
h)	Relación entre el ángulo sólido subtendido por la parte del fondo que satisface la presente especificación y el ángulo sólido subtendido por la imagen	≥ 9

2.4.2 Condiciones especiales

a)	Número típico de evaluadores por monitor a $4H$	2 (para la mitad de las sesiones) 3 (para la otra mitad)
b)	Número típico de evaluadores por monitor a $6H$	como anteriormente
c)	Monitor**	tamaño de pantalla de alta calidad, 22″-26″ (50 cm-60 cm)
d)	Brillo y contraste de la imagen	establecido vía PLUGE (véase la Recomendación 814)
e)	Número típico de evaluadores por monitor	5 (2 a $4H$ y 3 a $6H$ para la primera sesión, 3 a $4H$ y 2 a $6H$ para la sesión siguiente, y así sucesivamente)
f)	Tipo de sala(s) de observación	una sala, con tres paredes tapizadas de blanco y la cuarta (detrás) tapizada en gris.

2.5 Sesión de evaluación

Una sesión debería durar hasta media hora e incluir hasta unas 40 *presentaciones* (véase § 2.6).

Las sesiones se organizan por grupos de dos, para que todos los evaluadores puedan observar las imágenes o secuencias a $4H$ y $6H$. Si hay demasiadas condiciones de prueba para un solo par de sesiones, deberán organizarse pares suplementarios. Deberá utilizarse un orden aleatorio para las presentaciones (derivado, por ejemplo, de cuadrados grecolatinos); pero el orden de las condiciones de prueba debería disponerse de manera que los efectos sobre las evaluaciones del cansancio o de la adaptación se equilibren de una sesión a otra. Pueden repetirse algunas de las presentaciones en varias sesiones para comprobar su coherencia. Cada condición de prueba deberá mostrarse dos veces en la misma sesión.

---

* $6H$ es la distancia preferida para las evaluaciones de sistemas convencionales (625/50, 525/60), aunque también es aceptable la utilización de evaluadores a $4H$, siempre y cuando los resultados se den por separado o no se observe ninguna diferencia significativa en las medidas obtenidas.

** Cuando se utiliza más de una sala de observación, los monitores deberían igualarse cuidadosamente.

Las imágenes y degradaciones deberían presentarse en una secuencia seudoaleatoria y, preferentemente, en secuencias distintas para cada sesión. En cualquier caso, la misma imagen o secuencia de prueba no debe nunca presentarse en dos ocasiones sucesivas con los mismos niveles de degradación, o con niveles distintos.

La gama de degradaciones debería elegirse de manera que la mayoría de los observadores utilicen todas las notas; debería tratarse de obtener una nota media total (promedio de todas las apreciaciones emitidas durante el experimento) cercana a 3.

Una sesión no debe durar más de media hora aproximadamente, incluidas las explicaciones y los preliminares; asimismo la secuencia de prueba podría iniciarse con varias imágenes que indicasen la gama de degradaciones y las apreciaciones de esas imágenes no se tendrían en cuenta en los resultados finales.

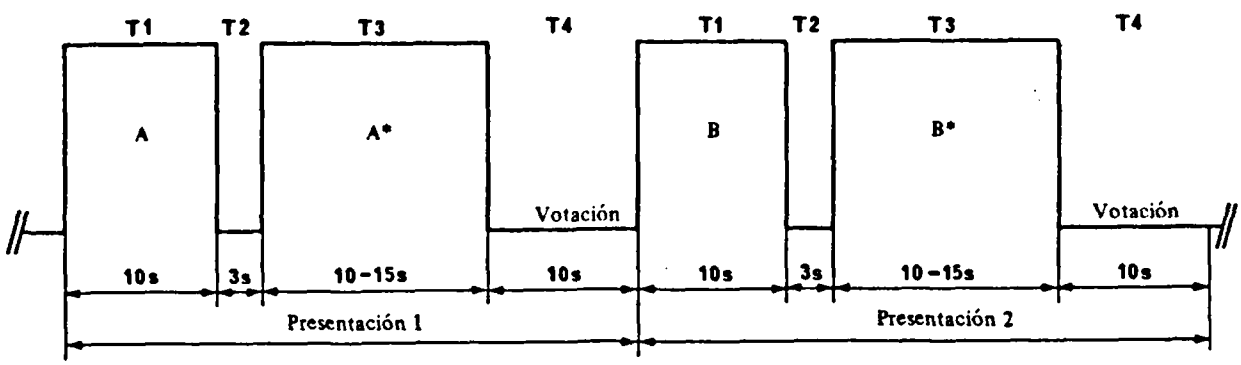
En el apéndice 2 se presentan otras ideas sobre la selección de niveles de degradaciones.

2.6 Presentación del material de prueba

Una sesión de prueba consta de varias presentaciones. La estructura de las presentaciones se indica en la fig. 2 siguiente.

FIGURA 2

Estructura de las presentaciones del material de prueba



A, B: Imagen o secuencia de referencia
A*, B*: Imagen o secuencia a evaluar

Cada presentación consta de cuatro fases:

T1	= 10 s	imagen de referencia
T2	= 3 s	gris mediano producido por un nivel vídeo de unos 200 mV
T3	= 10-15 s	condición a evaluar
T4	= 10 s	gris mediano

La duración de T3 puede ser de 10 a 15 s. Incluso para las imágenes en movimiento, resulta evidente que prolongar el periodo más allá de 15 s no mejora la capacidad del evaluador para juzgar las imágenes.

2.7 Observadores

En las pruebas participarán 15 observadores como mínimo. No han de ser expertos, en el sentido de que no estén directamente familiarizados con la calidad de imagen de televisión en su trabajo normal, ni tampoco sean evaluadores experimentados*. Antes de una sesión, debe examinarse a los observadores para determinar su agudeza visual normal o corregida, y su visión normal de colores, utilizando gráficos elegidos especialmente.

---

* Las conclusiones preliminares sugieren que, utilizando tecnologías de presentación y de calidad de transmisión más elevadas, los observadores no experimentados podrían dar lugar a resultados más críticos.

2.8 Escalas de apreciación

Debe utilizarse la escala de apreciación de cinco notas:

5 imperceptible

4 perceptible, pero no molesta

3 ligeramente molesta

2 molesta

1 muy molesta.

Los evaluadores deben utilizar un formulario que indique muy claramente la escala, y que cuente con cuadros numerados u otro medio para registrar las notas.

2.9 Selección del material de evaluación

Ciertos parámetros pueden dar lugar a un orden similar de degradaciones para la mayoría de las imágenes o secuencias. En esos casos, los resultados obtenidos con un número muy reducido de imágenes o secuencias (por ejemplo dos) pueden dar sin embargo una evaluación significativa.

Sin embargo, los nuevos sistemas a menudo tienen un impacto que depende mucho del contenido de la escena o de la secuencia. En esos casos, habrá una distribución estadística de la probabilidad de degradación y del contenido de la imagen o de la secuencia, para la totalidad de las horas de programa. Si, como es normal, no se conoce la forma de esa distribución, la selección de material de prueba y la interpretación de los resultados deben hacerse con sumo cuidado.

En general, es esencial incluir material crítico, porque se puede tener esto en cuenta cuando se interpretan los resultados, pero no es posible extrapolar a partir de material no crítico. En los casos en que el contenido de la escena o de la secuencia afecte a los resultados, deberá elegirse material que sea «crítico pero no indebidamente crítico» para el sistema sometido a prueba. La expresión «no indebidamente crítico» implica que las imágenes puedan formar parte, presumiblemente, de las horas normales de programación. En esos casos, deberían utilizarse por lo menos cuatro elementos, de los que la mitad sean absolutamente críticos, y la mitad moderadamente críticos.

Varias organizaciones han desarrollado imágenes fijas y secuencias de prueba. En el futuro se espera tratarlas en el marco del CCIR.

El CCIR ha propuesto material para evaluar sistemas digitales en que se aplica a señales de la Recomendación 601 una reducción de velocidad binaria a 30-33 Mbit/s. En la evaluación de esos sistemas ha de ser posible efectuar varias operaciones de procesamiento en puntos posteriores de la cadena, como la incrustación cromática. En esos casos, el sistema de incrustación debe incluirse en los trayectos de la señal por el sistema directo y por el sistema de prueba. Esas señales pueden incluirse entonces en las presentaciones de evaluación. Con este método es importante, empero, evitar imágenes o secuencias de referencia degradadas. Si interesa evaluar la deterioración adicional causada a una imagen ya degradada, ambas deberán utilizarse como secuencias de evaluación.

En los apéndices 1 y 2 se presentan otras ideas sobre la selección de materiales de prueba.

2.10 Introducción a las evaluaciones

Deberá familiarizarse cuidadosamente a los evaluadores con el método de evaluación, y con los tipos de degradación que probablemente se produzcan. Deberían permitirse preguntas para facilitar la compresión, pero no deben cambiarse las instrucciones de una sesión para otra, y se procurará la máxima coordinación en las respuestas.

Al principio de cada sesión, se darán explicaciones a los observadores sobre el tipo de evaluación, la escala de apreciación, la secuencia y la temporización (imagen de referencia, gris, imagen de evaluación, periodo de votación). La gama y el tipo de las degradaciones que van a evaluarse deberá ilustrarse con imágenes distintas de las utilizadas en las pruebas, pero de sensibilidad comparable. No debe darse a entender que la peor calidad observada corresponde necesariamente a la nota subjetiva más baja. Debe pedirse a los observadores que basen su apreciación en la impresión global que les da la imagen y que expresen esas apreciaciones en los mismos términos que se utilizan para definir la escala subjetiva.

Debe pedirse a los observadores que observen la imagen durante los periodos T1 y T3. La votación debe autorizarse únicamente durante T4.

2.11 Presentación de los resultados

Debe comprobarse la coherencia de los resultados examinando las notas dadas por el mismo observador a la misma imagen en la misma sesión. Si las evaluaciones difieren en dos o más notas, deben suprimirse ambos resultados.

Para cada parámetro de prueba debe darse la desviación media y típica de la distribución estadística de los grados de evaluación. Si lo que se evalúa es el cambio de degradación con un valor de parámetro variable, deben utilizarse técnicas de ajuste de curvas. El ajuste de curvas logístico y el eje logarítmico permitirán hacer una representación en línea recta, que es la forma de presentación preferida.

Los resultados deben darse junto con la información siguiente:

- detalles de la configuración del experimento,
- detalles de los materiales de evaluación,
- tipo de la imagen de origen y de los monitores,
- número y tipo de evaluadores,
- sistemas de referencias utilizados,
- nota media global del experimento,
- notas media original y ajustada, y desviaciones típicas si se ha eliminado uno o más observadores conforme al procedimiento detallado en § 3.

Luego de la sesión de prueba, se deben calcular los valores medios $E(X_j)$ y las desviaciones típicas $\sigma(x_j)$ asociadas a cada nivel de degradación o sistema de procesamiento en evaluación j. Estos valores medios están basados en una distribución, cuyas dos variables son las escenas y los observadores. Se debe examinar entonces si la distribución es normal o no lo es utilizando la prueba β_2 (por el cálculo del coeficiente de curtosis de la función, es decir, la razón entre el momento de cuarto orden y el cuadrado del momento de segundo orden). Si β_2 está comprendido entre 2 y 4, la distribución puede considerarse normal. Las notas $X_{i,j}$ de cada distribución j deben compararse entonces con el valor medio asociado más dos veces la desviación típica asociada (si es normal) o $\sqrt{20}$ veces (si no es normal), P_i, y el valor medio asociado menos dos veces la misma desviación típica o $\sqrt{20}$ veces, Q_i. Cada vez que una nota del observador sea superior o inferior a esta gama, deben ser registradas en un contador asociado a cada observador. Se han de utilizar dos contadores separados para valores superiores, P_i, e inferiores, Q_i. Por último, se deben calcular las dos relaciones siguientes: $P_i + Q_i$ sobre el número total de notas de cada observador durante la sesión entera, y $P_i - Q_i$ sobre $P_i + Q_i$ como valor absoluto. Si la primera es mayor del 5% y la última menor del 30%, se debe rechazar al observador i*.

Matemáticamente, dicho procedimiento puede expresarse también como sigue:

si $X_{i,j} \geq E(X_j) + 2 \cdot \sigma(X_j)$ (distribución normal)

o

si $X_{i,j} \geq E(X_j) + \sqrt{20} \cdot \sigma(X_j)$ (distribución no normal)

se obtiene $P_i = P_i + 1$.

Si $X_{i,j} \leq E(X_j) - 2 \cdot \sigma(X_j)$ (distribución normal)

o

si $X_{i,j} \leq E(X_j) - \sqrt{20} \cdot \sigma(X_j)$ (distribución no normal)

se obtiene $Q_i = Q_i + 1$.

$$\text{Si } \frac{P_i + Q_i}{\text{total de notas por observador}} > 0{,}05 \quad \text{y} \quad \left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0{,}3$$

en este caso, se rechaza el observador i.

---

* Este procedimiento no debe aplicarse más de una vez a los resultados de un experimento determinado. Además, el empleo del procedimiento ha de ser limitado a los casos en los que haya relativamente pocos observadores (por ejemplo, menos de 20), todos ellos no especializados.

3. El método de doble estímulo con escala de calidad continua

3.1 Descripción general

Una evaluación típica puede ser aplicable a la evaluación de un nuevo sistema o de los efectos de los trayectos de transmisión sobre la calidad. Se considera que el método de doble estímulo es especialmente útil cuando no se pueden proporcionar estímulos de prueba que abarquen toda la gama de calidad.

El método es cíclico puesto que se pide al evaluador que observe un par de imágenes, ambas de la misma fuente, pero habiéndose transmitido una por el sistema que se evalúa, y la otra directamente desde la fuente. Se le pide que evalúe la calidad de ambas.

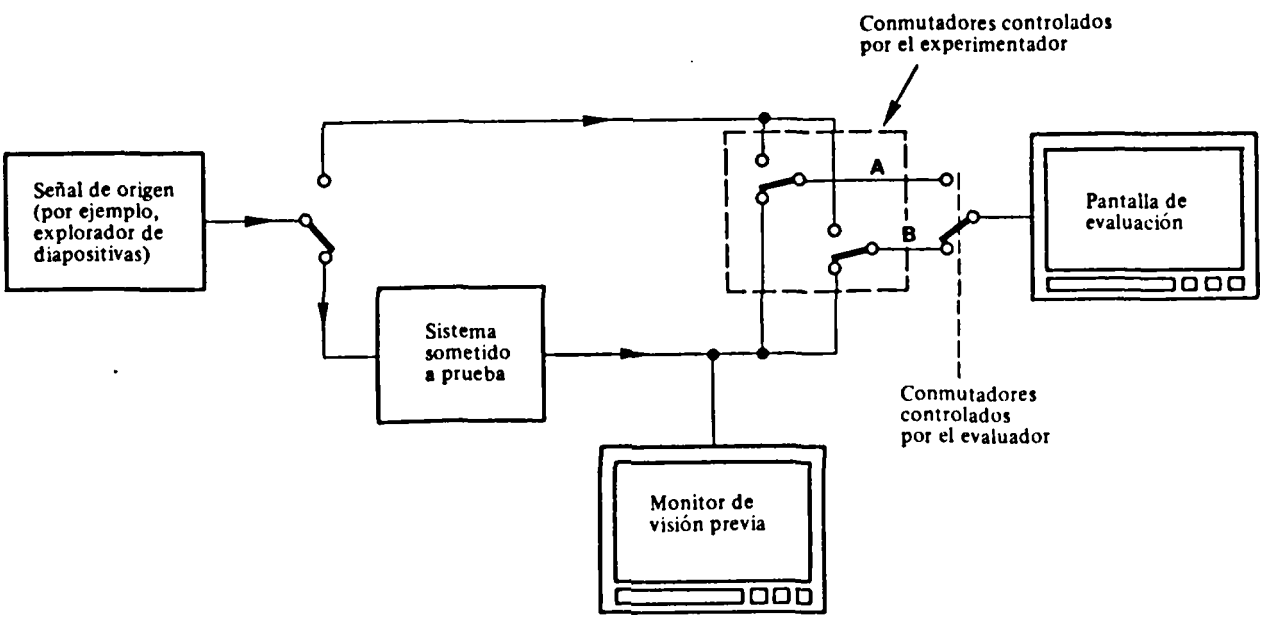
En sesiones que duran hasta media hora, se presenta al evaluador una serie de pares (aleatorios) de imágenes en orden aleatorio, y con degradaciones aleatorias que abarcan todas las combinaciones requeridas. Al final de las sesiones, se calculan las notas medias para cada condición de prueba y para cada imagen de prueba.

3.2 Disposición general

La disposición general del sistema de prueba debería ser la que se indica a continuación en la fig. 3.

FIGURA 3

Disposición general del sistema de prueba para el método de doble estímulo con escala de calidad continua



A continuación se indican dos variantes, (I) y (II), de este método.

(I) El evaluador, que suele estar solo, puede conmutar entre las dos condiciones A y B hasta que esté convencido de que se ha hecho una opinión de cada una. Las líneas A y B reciben la imagen directa de referencia, o la imagen transmitida por el sistema sometido a prueba, pero la transmisión por una línea u otra varía aleatoriamente entre una condición de prueba y la siguiente; el experimentador anota ese dato, pero no lo anuncia.

(II) Los evaluadores observan sucesivamente las imágenes de las líneas A y B, para hacerse una opinión de cada una. Las líneas A y B se alimentan para cada presentación de la misma manera que anteriormente (I). Todavía se está investigando la estabilidad de los resultados de esta variante con una gama limitada de calidad.

3.3 Señales de origen

Para este método son también válidas las consideraciones que se han hecho para el método del § 2. Sin embargo, una referencia degradada puede no tener el mismo efecto sobre la estabilidad.

3.4 Condiciones de observación

También aquí son válidas las consideraciones del método del § 2. Sin embargo, para la variante (I), el número de evaluadores por monitor es 1.

3.5 Sesión de evaluación

Como para el método del § 2. Para la variante (I) por lo menos, no es necesario tratar de obtener una nota media global de 3.

3.6 Presentación del material de evaluación

Una sesión de evaluación consta de varias presentaciones. En la variante (I), que tiene un solo observador, el evaluador puede conmutar libremente entre las señales A y B para cada presentación, hasta que tenga la medida mental de la calidad asociada con cada señal. Puede, por ejemplo, decidir hacerlo en dos o tres veces por periodos de hasta 10 s. En la variante (II), que utiliza simultáneamente varios observadores, antes de registrar los resultados, se muestra el par de condiciones una o más veces durante un lapso de tiempo similar, para permitir al evaluador adquirir la medida mental de las calidades asociadas con éstas; a continuación, cada par de condiciones se presenta nuevamente una o más veces, mientras se registran los resultados. El número de repeticiones depende de la duración de las secuencias de prueba. Para las imágenes fijas, puede ser apropiada una secuencia de 3-4 s y cinco repeticiones (votándose en las dos últimas). Para imágenes en movimiento con efectos secundarios variables en el tiempo, parece adecuada una secuencia de 10 s, con dos repeticiones (votándose en la segunda).

Cuando consideraciones de índole práctica limitan la duración de las secuencias disponibles a menos de 10 s, pueden efectuarse composiciones utilizando estas secuencias más breves como segmentos, para ampliar el tiempo de exhibición a 10 s. Con el objeto de reducir a un mínimo la discontinuidad en los empalmes, los segmentos de secuencias sucesivas pueden ser invertidos en el tiempo (lo que se denomina, a veces exhibición «palindrómica»). Conviene asegurarse de que las condiciones de prueba exhibidas como segmentos invertidos en el tiempo representen procesos causales, es decir, deben ser obtenidos haciendo pasar la señal de origen invertida en el tiempo a través del sistema que se está probando.

3.7 Observadores

Como para el método del § 2.

3.8 Escala de apreciación

El método requiere la evaluación de dos versiones de cada imagen de prueba. Una de las imágenes de prueba de cada par está degradada mientras que la otra puede o no contener una degradación. La imagen no degradada se incluye como referencia, pero no se dice a los observadores cuál es la imagen de referencia. En las series de pruebas, se cambia la posición de la imagen de referencia, de manera seudoaleatoria.

Se pide simplemente a los observadores que evalúen la calidad global de imagen de cada presentación haciendo una marca en una escala vertical. Las escalas verticales se imprimen por pares para respetar la presentación doble de cada imagen de prueba. Las escalas ofrecen un sistema de evaluación continuo para evitar errores de cuantificación, pero están divididas en cinco segmentos de igual longitud que corresponden a la escala de calidad normal de cinco notas del CCIR. Los términos asociados que distinguen los distintos niveles son los mismos que se utilizan normalmente, pero en este caso se incluyen como indicación, y se imprimen solamente en el lado izquierdo de la primera escala de cada línea de diez columnas dobles en la hoja de resultados. En la fig. 4 se muestra una sección de una hoja típica de resultados. Las posibilidades de confusión entre las divisiones de la escala y los resultados de prueba se evitan imprimiendo las escalas en azul y registrando los resultados en negro.

3.9 Selección del material de evaluación

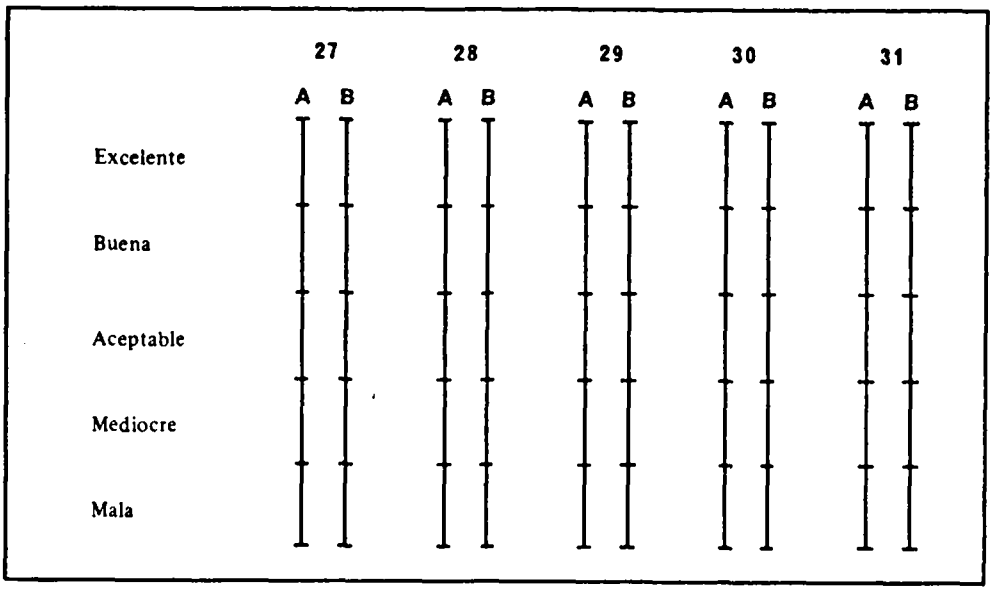
Como para el método del § 2.

3.10 Introducción a la evaluación

Como para el método del § 2, excepto para el último párrafo del § 2.10.

FIGURA 4

Parte de una hoja de evaluación de calidad en que se utilizan escalas continuas



3.11 Presentación de los resultados

Los resultados se pueden presentar de dos maneras distintas:

- En primer lugar, los resultados pueden expresarse en forma de pruebas comparativas, es decir para que indiquen directamente el cambio de calidad con respecto a la condición de referencia. Para cada parámetro de prueba deben darse la media y la desviación típica de la distribución estadística de la diferencia medida.

- En segundo lugar (método de presentación preferido), los resultados pueden convertirse en los términos utilizados para describir un grado de calidad equivalente. Los pares de evaluaciones (referencia y prueba) para cada condición de prueba se convierten de mediciones de longitud en la hoja de resultados a resultados normalizados en la escala de 0 a 100. Para cada sistema sometido a prueba, se promedian a continuación los resultados para los distintos grupos de observadores, diversas distancias de observación e imágenes de prueba, a fin de dar notas medias a las condiciones de referencia y de prueba para cada combinación de variables.

Como las notas medias de las condiciones de referencia son siempre inferiores a 1,0, debe cambiarse la escala de los resultados de las pruebas sustrayendo la degradación residual. La nota media para la condición de referencia se trata como degradación residual. Los resultados de la resta se expresan en unidades de degradación (imps) pero pueden volverse a traducir a notas medias si se desea.

El Informe debe aportar la misma información adicional que el método del § 2, excepto para la nota media.

4. Otros métodos de evaluación

En circunstancias apropiadas se deberían utilizar los métodos de estímulo único y de comparación de estímulos.

4.1 Métodos de estímulo único

En los métodos de estímulo único, se presenta un sola imagen o secuencia de imágenes y el evaluador da un índice de toda la presentación.

4.1.1 Observadores

Para las pruebas de laboratorio, los observadores se suelen seleccionar como en el § 2.7. El número necesario de evaluadores depende de la sensibilidad y de la fiabilidad del procedimiento de prueba adoptado y del tamaño previsto del efecto que se busca. En circunstancias normales, se utiliza una muestra de 10 a 20 evaluadores por prueba.

4.1.2 Imágenes de prueba

Para las pruebas de laboratorio debe seleccionarse el contenido de las imágenes de prueba como se describe en el § 2.9.

Una vez seleccionado el contenido, las imágenes de prueba se preparan para que reflejen las opciones de diseño estudiadas por la gama o gamas de uno o más factores. Cuando se examinan dos o más factores, las imágenes pueden prepararse de dos maneras: en la primera, cada imagen representa solamente un nivel de un factor, y en la segunda, cada imagen representa un nivel de cada factor examinado pero a lo largo de las imágenes se observa el nivel de cada factor con cada nivel de todos los demás factores. Ambos métodos permiten atribuir claramente resultados a efectos específicos. El segundo método permite también detectar las interacciones entre factores (es decir, los efectos no aditivos).

4.1.3 Condiciones de observación

Se ha notado que cuando los observadores actúan libremente pueden elegir distancias de observación mayores que las utilizadas en evaluaciones subjetivas. La relación entre las distancias de observación preferidas y las utilizadas en evaluaciones requieren ulterior estudio.

4.1.4 Sesión de evaluación

Antes de la sesión de evaluación se da a los observadores una descripción de la labor de observación y, normalmente, ejemplos de las imágenes o de las secuencias de imágenes. Las instrucciones suelen ser escritas o grabadas. Se procura no influenciar a los observadores en el cumplimiento de su tarea.

La sesión consiste en una serie de pruebas de evaluación, que deberían presentarse en secuencia aleatoria y, preferiblemente, en una secuencia aleatoria distinta para cada observador. Cuando se utiliza una secuencia aleatoria única, el experimentador normalmente se asegura de que la misma imagen no se presente dos veces seguidas con el mismo tipo y nivel de degradación.

Una prueba de evaluación típica consiste en tres presentaciones: un campo de adaptación en gris medio, un campo de estímulo, y un campo de post-exposición en gris medio. Las duraciones de esas presentaciones varían según la tarea del observador, los materiales (por ejemplo, móvil o inmóvil), y las opciones o factores examinados, no obstante duraciones de 3, 10 y 10 s respectivamente son bastante frecuentes. El índice o los índices del observador pueden recogerse durante la presentación del estímulo o del campo de post-exposición.

4.1.5 Tipos de métodos de estímulo único

En general, se han utilizado tres tipos de métodos de estímulo único en las evaluaciones de televisión.

4.1.5.1 Métodos de apreciación por categorías

En las apreciaciones por categorías, los observadores asignan una imagen o secuencia de imágenes a una categoría elegida entre un conjunto de categorías que, por lo general, se definen en términos semánticos. Las categorías pueden reflejar apreciaciones, o si se detecta o no un atributo (por ejemplo, para establecer el umbral de degradación). Las escalas de categorías que evalúan la calidad de imagen y la degradación de imagen, son las que se han utilizado más a menudo; las escalas del CCIR se dan en el cuadro 4 siguiente. En controles operacionales se utilizan a veces medias notas. Las escalas que evalúan la legilibilidad del texto, el esfuerzo de lectura, y la utilidad de la imagen se han utilizado en casos especiales.

Este método permite distribuir las apreciaciones en una escala de categorías para cada condición. El análisis de las respuestas depende de la apreciación (detección, etc.) y de la información buscada (umbral de detección, rangos o tendencia media de las condiciones, «diferencias» psicológicas entre condiciones). Se dispone de numerosos métodos de análisis.

CUADRO 4

Escalas de calidad y degradación del CCIR

Escala de cinco notas	
Calidad	Degradación
5 Excelente	5 Imperceptible
4 Buena	4 Perceptible, pero no molesta
3 Aceptable	3 Ligeramente molesta
2 Mediocre	2 Molesta
1 Mala	1 Muy molesta

4.1.5.2 Métodos que no utilizan una escala de evaluación por categorías

Cuando las apreciaciones no se hacen por categorías, los observadores asignan un valor a cada imagen o secuencia de imagen mostrada. Este método puede revestir las dos formas siguientes:

En la apreciación por escala continua, variante del método por categorías, el evaluador asigna cada imagen o secuencia de imagen a un punto de una línea trazada entre dos niveles semánticos (por ejemplo, los valores extremos de una escala de categorías como la del cuadro 4). La escala puede incluir rangos adicionales en puntos intermedios para fines de referencia. La distancia con respecto a un extremo de la escala se toma como índice para cada condición.

En la distribución por escala mumérica, el evaluador asigna a cada imagen o secuencia de imágenes un número que refleja su nivel estimado en una dimensión especificada (por ejemplo, nitidez de la imagen). La escala de números utilizada puede ser restringida (por ejemplo, 0 a 100) o no. A veces, el número asignado describe el nivel juzgado en términos «absolutos» (sin ninguna relación directa con el nivel de cualquier otra imagen o secuencia de imágenes, como en ciertas formas de estimaciones de magnitud). En otros casos, el número describe el nivel juzgado en relación al de un «estándar» visto anteriormente (por ejemplo, estimación de magnitud, fraccionamiento, y estimación de relación).

Con ambas formas se obtiene una distribución de números para cada condición. El método de análisis utilizado depende de la naturaleza de la apreciación y de información requerida (por ejemplo, rangos, tendencia media, «diferencias» psicológicas).

4.1.5.3 Métodos de realización

Ciertos aspectos de la observación normal pueden expresarse como realización de tareas concretas (hallar una información determinada, leer un texto, identificar objetos, etc.). Así pues, como índice de la imagen o secuencia de imágenes puede utilizarse una medida de realización (por ejemplo, la precisión o velocidad con que se realizan esas tareas).

Los métodos de realización llevan a distribuciones de notas de precisión o de velocidad para cada condición. El análisis trata sobre todo de establecer relaciones entre las condiciones de la tendencia media (y dispersión) de las notas, y a menudo utiliza el análisis de varianza o una técnica similar.

4.1.6 Cuestiones

4.1.6.1 Gama de condiciones y anclaje

Dado que el método por categorías y otros métodos son sensibles a las variaciones de la gama y de la distribución de las condiciones observadas, las sesiones de evaluación deberían incluir las gamas completas de los factores sometidos a variación. Sin embargo, puede hacerse una aproximación con una gama más restringida, presentando también ciertas condiciones que se situarían en los extremos de las escalas. Podrían representarse esas condiciones como ejemplo, e identificarlas como las más extremas (anclaje directo), o distribuirlas en la sesión y no identificarlas como más extremas (anclaje indirecto).

4.1.6.2 Significado de los resultados

Como varían con la gama, puede ser inadecuado interpretar las apreciaciones a partir del método por categorías y de otros métodos en términos absolutos (por ejemplo, la calidad de una imagen o secuencia de imágenes).

4.2 Métodos de comparación de estímulos

En los métodos de comparación de estímulos, se presentan en pantalla dos imágenes o secuencias de imágenes y el observador da un índice de la *relación* entre las dos presentaciones.

4.2.1 Evaluadores

La determinación de los evaluadores se lleva a cabo de la misma manera que en los métodos de estímulo único.

4.2.2 Imágenes de prueba

Las imágenes o secuencias de imágenes utilizadas se generan de la misma manera que en los métodos de estímulo único. Las imágenes o secuencias de imágenes resultantes se combinan entonces para constituir los pares que se utilizan en las pruebas de evaluación.

4.2.3 Condiciones de observación

Las condiciones de observación se determinan de la misma manera que en los métodos de estímulo único.

4.2.4 Sesión de evaluación

En la prueba de evaluación se utilizará un monitor, o bien dos monitores debidamente sincronizados, y se procederá en general como en los casos de estímulos únicos. Con un solo monitor, se utilizarán dos campos de estímulos idénticos. En ese caso, conviene que, en la distintas pruebas, ambos miembros de un par aparezcan el mismo número de veces en primera y en segunda posición. Si se utilizan dos monitores, los campos de estímulos se muestran simultáneamente.

4.2.5 Tipos de métodos de comparación de estímulos

En las evaluaciones de televisión se han utilizado los tres tipos de métodos de comparación de estímulos que figuran a continuación.

4.2.5.1 Métodos de apreciación por categorías

En los métodos de apreciación por categorías, los observadores asignan la relación entre miembros de un par a una categoría elegida entre un conjunto de categorías que, normalmente, se definen en términos semánticos. Esas categorías pueden indicar la existencia de diferencias perceptibles (por ejemplo, IGUAL, DIFERENTE), la existencia y dirección de diferencias perceptibles (por ejemplo, MENOS, IGUAL, MÁS), o apreciaciones de amplitud y dirección. La escala de comparación del CCIR se indica en el cuadro 5 siguiente.

CUADRO 5

Escala de comparación

–3	Mucho peor
–2	Peor
–1	Ligeramente peor
0	Igual
+1	Ligeramente mejor
+2	Mejor
+3	Mucho mejor

Este método proporciona una distribución de las apreciaciones en categorías de escalas para cada par de condiciones. La manera en que se analizan las respuestas depende de la apreciación (por ejemplo, diferencia) y de la información requerida (por ejemplo, diferencias apenas perceptibles, rangos de condiciones, «diferencias» entre condiciones, etc.).

4.2.5.2 Métodos que no utilizan una escala de apreciación por categorías

Cuando las apreciaciones no se hacen por categorías, los observadores asignan un valor a la relación entre los elementos de un par de evaluación. Este método puede revestir dos formas:

- En la apreciación con escala continua, el evaluador asigna cada relación a un punto de una línea trazada entre dos notas (por ejemplo, IGUAL-DIFERENTE, o los extremos de una escala por categorías como en el cuadro 5). Las escalas pueden incluir marcas de referencia adicionales en puntos intermedios. La distancia con respecto a un extremo de la línea se toma como valor para cada par de condiciones.

- En la segunda forma, el evaluador asigna a cada relación un número que refleja el nivel estimado en una dimensión especificada (por ejemplo, diferencia de calidad). La gama de números utilizada puede ser limitada o no. El número asignado puede describir la relación en términos «absolutos» o en términos de la relación en un par «estándar».

Con ambas formas se obtiene una distribución de valores para cada par de condiciones. El método de análisis depende de la naturaleza de la apreciación y de la información requerida.

4.2.6 Métodos de realización

En algunos casos, las mediciones de realización pueden derivarse de procedimientos de comparación de estímulos. En el método de elección forzada, el par se dispone para que un elemento contenga un nivel particular de un atributo (por ejemplo, degradación), mientras que el otro contiene un nivel diferente o ninguno de ese atributo. Se pide al observador que decida qué elemento contiene el mayor o menor nivel del atributo o cuál contiene algo del atributo; la precisión y la velocidad de la realización se toman como índices de la relación entre los miembros del par.

4.2.7 Cuestiones

4.2.7.1 Formación de los pares

Los métodos de comparación de estímulos determinan más completamente las relaciones entre condiciones cuando en las apreciaciones se comparan todos los pares posibles de condiciones. Sin embargo, si esto requiere un número excesivo de observaciones, éstas podrían dividirse entre los evaluadores, o podría utilizarse una muestra de todos los pares posibles.

4.2.7.2 Métodos con escala multidimensional

Varios investigadores han utilizado métodos con escala multidimensional para estudiar las apreciaciones de comparación de estímulos de televisión.

4.3 Selección de los métodos

Todos los métodos descritos hasta ahora tienen sus ventajas y sus limitaciones, y todavía no es posible recomendar uno preferentemente. Por consiguiente, la selección del método más apropiado a las circunstancias se deja al buen criterio del investigador.

Las limitaciones de los diversos métodos sugieren que podría no ser acertado dar demasiada importancia a un solo método, por lo que convendría estudiar planteamientos más «completos» como la utilización de varios métodos o un planteamiento multidimensional.

APÉNDICE 1
AL ANEXO 1

Características de fallo del contenido de la imagen 1

1. Introducción

Luego de su realización, un sistema estará sujeto a una gama potencialmente amplia de material de programa, alguno del cual podría no hallar el modo de tener cabida sin pérdida de calidad. Al considerar la aptitud de un sistema es necesario conocer la proporción de material de programa que resultará crítico para el sistema y la pérdida de calidad que se aguarda en tales casos. En efecto, es necesario disponer de la característica de fallo del contenido de la imagen para el sistema en estudio.

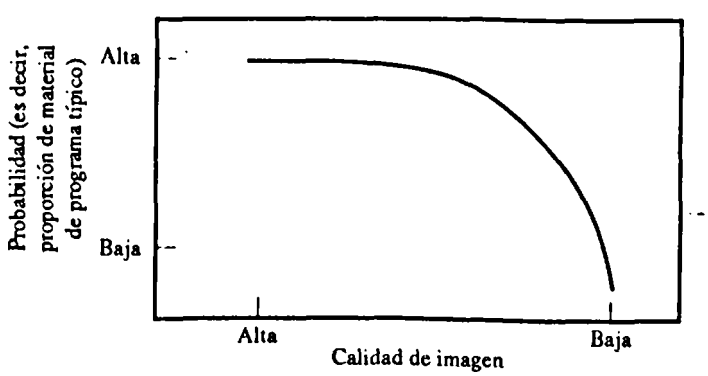
Dicha característica de fallo es particularmente importante para sistemas cuya calidad de funcionamiento puede no degradarse uniformemente a medida que el material se torna cada vez más crítico. Por ejemplo, ciertos sistemas digitales y adaptables pueden mantener un alto grado de calidad sobre una amplia gama de material de programa, pero se degradan fuera de ésta.

2. Obtención de la característica de fallo

En términos conceptuales, una característica de contenido de imagen determina la proporción de material para la que a largo plazo es probable que el sistema alcance niveles particulares de calidad. Este concepto se ilustra en la fig 5.

FIGURA 5

Representación gráfica de una característica de fallo de contenido de imagen posible



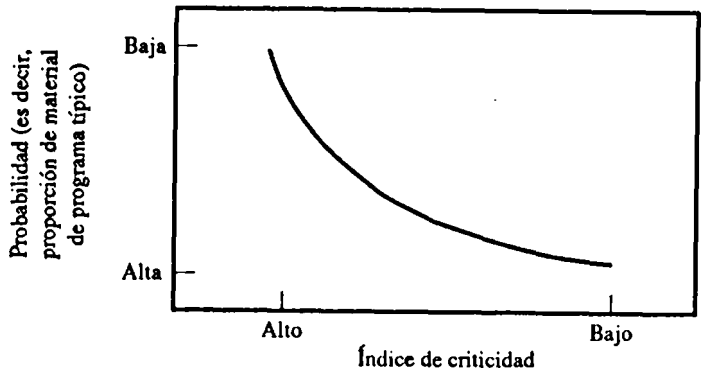
Una característica de fallo de contenido de imagen puede obtenerse en cuatro pasos:

Paso 1 – Determinación de una medida algorítmica de «criticidad» que fuera capaz de clasificar un número de secuencias de imagen que han estado sometidas a distorsión proveniente del sistema o clases de sistemas afectados, de manera tal que la categoría de clasificación corresponda a la que se obtendría si la tarea se hubiera efectuado por medio de observadores. Esta medida de criticidad puede implicar aspectos de modelado visual.

Paso 2 – Obtención, por aplicación de la medida de criticidad a un gran número de muestras tomadas de la televisión típica, de una distribución que estima la probabilidad de ocurrencia de material que proporciona distintos niveles de criticidad para el sistema, o clases de sistemas en estudio. En la fig. 6 se ilustra un ejemplo de dicha distribución.

FIGURA 6

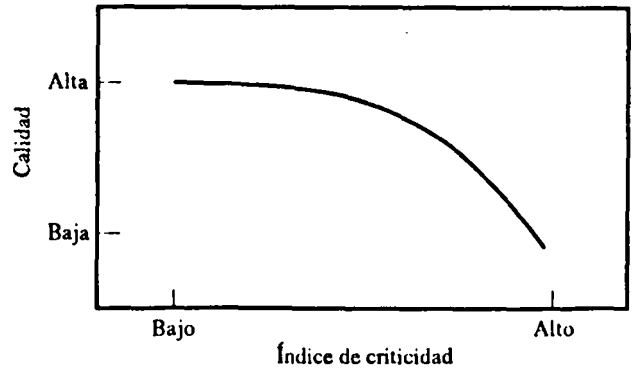
Probabilidad de ocurrencia de material de niveles de criticidad diferentes



Paso 3 – Obtención, por medios empíricos, de la capacidad del sistema para mantener la calidad a medida que aumenta el nivel de criticidad. En la práctica, esto requiere la evaluación subjetiva de la calidad alcanzada por el sistema con material seleccionado para muestrear el margen de criticidad identificado en el paso 2. Esto da por resultado una función que relaciona la calidad alcanzada por el sistema y el nivel de criticidad en material de programa. En la fig. 7 se ilustra un ejemplo de dicha función.

FIGURA 7

Una función posible que relaciona la calidad con la criticidad del material de programa



Paso 4 – Conlleva la información de los pasos 2 y 3 a fin de obtener una característica de fallo de contenido de imagen de la forma indicada en la fig. 5.

3. Utilización de la característica de fallo

La característica de fallo, que proporciona una imagen de la calidad de funcionamiento que probablemente se obtenga a través de la gama de material de programa posible, constituye un instrumento importante para considerar la adaptabilidad de los sistemas. La característica de fallo se puede utilizar de tres maneras:

- para optimizar parámetros (por ejemplo, resolución de la fuente, velocidad binaria, anchura de banda) de un sistema en la etapa de diseño, para adaptarlo más estrechamente a las necesidades de un servicio;
- para estudiar la adecuación de un sistema (es decir, anticipar la incidencia y gravedad del fallo durante la operación);
- para evaluar las adecuaciones relativas de sistemas de alternativa (es decir, comparar las características de fallo y determinar qué sistema sería más adecuado para el uso). Cabe señalar que, mientras que los sistemas de alternativa de tipo semejante pueden utilizar el mismo índice de criticidad, es posible que los sistemas de tipo no semejante puedan tener distintos índices de criticidad. Sin embargo, como la característica de fallo sólo expresa la probabilidad de que en la práctica se vean diferentes niveles de calidad, las características se pueden comparar directamente aun cuando provengan de índices de criticidad de sistemas específicos diferentes.

Si bien el método descrito en la presente Recomendación proporciona un medio para medir la característica de fallo de contenido de imagen de un sistema, no podría utilizarse para predecir totalmente la aceptabilidad del sistema por el espectador de un servicio de televisión. Para obtener esta información puede ser necesario que una cantidad de telespectadores vean programas codificados con el sistema de interés, y estudiar luego sus comentarios.

APÉNDICE 2
AL ANEXO 1

Método para determinar una característica de fallo compuesta para contenido de programa y condiciones de transmisión

1. Introducción

Una característica de fallo compuesta relaciona la calidad de imagen percibida con la probabilidad de ocurrencia en la práctica de una forma tal que considere explícitamente el contenido de programa y las condiciones de transmisión.

En principio, dicha característica se podría obtener por medio de un estudio subjetivo que exige una cantidad suficiente de observaciones, momentos de prueba y puntos de recepción para producir una muestra que represente la población de contenido de programa y condiciones de transmisión posibles. Sin embargo, en la práctica, un experimento de este tipo sería irrealizable.

En el presente apéndice se describe un procedimiento alternativo, más fácilmente realizable, para determinar las características de fallo compuestas. Este método consta de tres etapas:

- análisis del contenido de programa;
- análisis del canal de transmisión; y
- obtención de las características de fallo compuestas.

2. Análisis del contenido de programa

Esta etapa exige dos operaciones: primero, se obtiene una medida apropiada del contenido del programa; y, segundo, se estiman las probabilidades con las que los valores de esta medición ocurren en la práctica.

La medición del contenido de programa es una estadística que recoge aspectos del contenido de programa que acentúan la capacidad del sistema(s) en estudio para proporcionar reproducciones fieles de material de programa desde el punto de vista perceptivo. Evidentemente, sería ventajoso que estuviera basada en un modelo de percepción apropiado. Sin embargo, en ausencia de tal modelo, podría ser suficiente una medición que recogiera algún aspecto de la diversidad espacial sobre tramas/cuadros de vídeo, siempre que esta medición presente una relación aproximadamente monótona con la calidad de la imagen percibida. Podría ser necesario utilizar diferentes mediciones para sistemas (o clases de sistemas) que emplean planteamientos fundamentalmente distintos para la representación de la imagen.

Una vez escogida la medición apropiada, es necesario estimar las probabilidades con las que los posibles valores de esta estadística ocurren. Esto se puede efectuar en una de las dos maneras siguientes:

- con el procedimiento empírico, en el que se realiza una muestra tomada al azar de unos doscientos segmentos de programa de 10 s en un formato de estudio adecuado en resolución, frecuencia de cuadro, y relación dimensional de la imagen al sistema(s) considerado. El análisis de esta muestra revela que para valores de la estadística que en la práctica se toman como estimaciones de probabilidad de ocurrencia se producen relativas frecuencias de ocurrencia; o
- con el método teórico, por el que se utiliza un modelo teórico para estimar las probabilidades. Se hace notar que, aunque se prefiere el método empírico, puede ser necesario en determinados casos emplear el método teórico (por ejemplo, cuando no se dispone de suficiente información sobre el contenido de programa, tal como la aparición de nuevas tecnologías de producción).

Los análisis precedentes darán por resultado una distribución de probabilidad para valores de la estadística de contenido (véase también el apéndice 1). Esto se combinará con los resultados del análisis de las condiciones de transmisión para preparar la etapa final del proceso.

3. Análisis del canal de transmisión

Esta etapa también exige dos operaciones: primero, se obtiene una medición de la calidad de funcionamiento del canal de transmisión; y, segundo, se estiman las probabilidades con las que los valores de esta medición ocurren en la práctica.

La medición de un canal de transmisión es una estadística que recoge aspectos de la calidad de funcionamiento de un canal que influencia la capacidad del sistema(s) en estudio para proporcionar reproducciones fieles de material fuente desde el punto de vista perceptivo. Evidentemente, sería ventajoso que esta medida se basara en un modelo de percepción apropiado. Sin embargo, en ausencia de tal modelo, sería suficiente una medida que recoja en cierto grado el stress impuesto por el canal, siempre que esta medida presente una relación aproximadamente monótona con la calidad de la imagen percibida. Puede ser necesario utilizar diferentes medidas para sistemas (o clases de sistemas) que emplean enfoques esencialmente distintos para la codificación del canal.

Una vez seleccionada la medida apropiada, es necesario estimar las probabilidades con las que los valores posibles de esta estadística ocurren. Esto puede efectuarse en una de las dos maneras siguientes:

- con el procedimiento empírico, en el que se mide la calidad de funcionamiento del canal en unos 200 momentos y puntos de recepción seleccionados al azar. El análisis de esta muestra revela funciones de ocurrencia relativas para valores de la estadística que se toman como estimación de probabilidad de ocurrencia en la práctica; o

- con el método teórico, en el que se utiliza un modelo teórico para estimar las probabilidades. Se hace notar que, aunque se prefiere el método empírico, puede ser necesario en determinados casos emplear el método teórico (por ejemplo, cuando no se dispone de suficiente información acerca de la calidad de funcionamiento del canal, tal como la aparición de nuevas tecnologías de transmisión).

Los análisis precedentes darán por resultado una distribución de probabilidad para valores de la estadística de canal. Esto se combinará con los resultados del análisis de contenido de programa para preparar la etapa final del proceso.

4. Obtención de las características de fallo compuestas

Esta etapa incluye un experimento subjetivo en el cual el contenido de programa y las condiciones de transmisión se varían conjuntamente de acuerdo con las probabilidades establecidas en las primeras dos etapas.

El método básico utilizado es el procedimiento de doble estímulo con escala de calidad continua y, en particular, la versión recomendada de 10 s para secuencias en movimiento (véase el anexo 1, § 3). Aquí, la referencia es una imagen con calidad de estudio en un formato apropiado (por ejemplo, un formato con resolución, frecuencia de trama, formato de imagen apropiado al sistema(s) en estudio). En contraste, la prueba presenta la misma imagen como si hubiera sido recibida por el sistema(s) en estudio bajo condiciones de canal seleccionado.

El material de prueba y las condiciones de canal se seleccionan de acuerdo con las probabilidades establecidas en las primeras dos etapas del presente método. Los segmentos del material de prueba, analizados cada uno de ellos para determinar su valor predominante de acuerdo con la estadística de contenido, incluyen un fondo común de selección. El material se muestra entonces a partir de este formato común de modo tal que abarca la gama de valores posibles de la estadística, escasamente en niveles menos críticos y más densamente en niveles más críticos. Los valores posibles de la estadística de canal se seleccionan en forma similar. Luego, estas dos fuentes de influencia independientes se combinan al azar para producir condiciones de canal contenido combinado de probabilidad conocida.

Los resultados de tales estudios, que relacionan la calidad de la imagen percibida con la probabilidad de ocurrencia en la práctica, se utilizan entonces para estudiar la adecuación de un sistema o comparar sistemas en términos de adecuación.

ANEXO 2

Métodos de evaluación de la calidad de la imagen en relacion con las degradaciones debidas a la codificación digital de las señales de televisión

1. Introducción

En el anexo 1 se indican métodos subjetivos para evaluar la calidad de imagen en la televisión con resolución convencional así como su degradación. Para TVAD se indican en la Recomendación 710. En el presente anexo se analiza la aplicación de esos métodos a la evaluación de los códecs de televisión.

Ultimamente, se ha adquirido una gran experiencia en lo relativo a la evaluación de la calidad de funcionamiento de códecs de alta calidad para televisión de componentes con relación 4:2:2 a 34, 45 y 140 Mbit/s. En las correspondientes pruebas, se examinó la calidad de funcionamiento de los códecs en términos de calidad de imagen decodificada básica, calidad después del tratamiento posterior en estudio (incrustación cromática y cámara lenta) aplicado a las imágenes decodificadas y la degradación de la imagen decodificada, asociada con la presencia de una gama de proporciones de bits erróneos en el canal. Algunas partes de este anexo se benefician de esas pruebas.

Las especificaciones de calidad en el caso de aplicaciones de distribución pueden expresarse en términos de la apreciación subjetiva de los observadores. En teoría, por tanto, esos códecs pueden evaluarse subjetivamente, contrastándolos con estas especificaciones. Sin embargo, la calidad de un códec diseñado para aplicaciones de contribución no podría especificarse teóricamente en términos de parámetros subjetivos de calidad de funcionamiento, porque su salida no está destinada a una visualización inmediata sino a tratamiento posterior en estudio, almacenamiento y/o codificación para transmisión ulterior. Dada la dificultad de definir esa calidad de funcionamiento para una diversidad de operaciones de tratamiento posterior, el enfoque preferido ha sido especificar la calidad de funcionamiento de una cadena de equipo, incluyendo una función de tratamiento posterior, a la que se considera representativa de una aplicación práctica de contribución. Esta cadena podría constar típicamente de un códec, seguido por una función de tratamiento posterior de estudio (o de otro códec en el caso de evaluación de calidad de contribución básica) seguido todavía por otro códec antes de que la señal alcance al observador. La adopción de esta estrategia para las especificaciones de códecs destinados a aplicaciones de contribución significa que los procedimientos de medición que se dan en la presente Recomendación pueden también utilizarse para su evaluación.

A lo largo del presente anexo se insiste en la importancia de elegir secuencias de imágenes de prueba críticas, sobre todo de escenas naturales, y se dan algunas directrices sobre cómo generar o escoger tales secuencias.

2. Evaluación subjetiva de la calidad de imagen de los códecs

Aunque se esté progresando al respecto, en la actualidad no se dispone de suficiente experiencia para dar detalles sobre métodos de evaluación objetiva de la calidad de imagen de los códecs. En materia de evaluación subjetiva, de la que existe mucha experiencia, se pueden hacer recomendaciones sobre condiciones de prueba y metodologías. Debe recordarse no obstante, al especificar objetivos de calidad o degradación, que los métodos existentes no pueden dar valoraciones subjetivas absolutas sino más bien resultados que están influidos en cierta medida por la elección de las condiciones de referencia y/o fijación. Pueden adoptarse las mismas metodologías para códecs de longitud de palabra fija y variable y para códecs de intratrama e intercuadro, aunque la elección de las secuencias de imágenes de prueba puede verse influenciada.

El método de evaluación más fiable para establecer un orden de jerarquía para los códecs de gran calidad consiste, en la actualidad, en evaluar todos los sistemas presentados al mismo tiempo y en condiciones idénticas. Las pruebas hechas independientemente, en las que se evalúan diferencias de calidad muy pequeñas, deben servir de guía más bien que de evidencia incuestionable de superioridad.

2.1 Evaluación de la calidad básica

Cuando se evalúa un códec para aplicaciones de distribución, esta calidad se refiere a las imágenes decodificadas después de un paso único a través de un par de códecs. En el caso de códecs de contribución, puede evaluarse la calidad básica después de varios códecs en serie, con el fin de simular así una aplicación típica de contribución.

2.1.1 Condiciones de observación y elección de los observadores

Se recomienda que las condiciones de observación y elección de observadores se efectúe de conformidad con el § 2.4 del anexo 1 a esta Recomendación para televisión con resolución convencional y según la Recomendación 710 en el caso de códecs de TVAD.

2.1.2 Utilización de secuencias de imágenes de prueba

Se recomienda que en la evaluación se utilicen secuencias de al menos seis imágenes, más una adicional para los efectos de demostración antes del comienzo de la prueba. Las secuencias podrán tener una duración del orden de 10 s, pero debe señalarse que los evaluadores pueden preferir una duración de 15 a 30 s. Deben variar entre moderadamente críticas y críticas en el contexto de la aplicación de reducción de velocidad binaria que esté en consideración.

2.1.3 Metodología de la prueba

Cuando la gama de calidades por evaluar es pequeña, lo que ocurrirá normalmente en el caso de códecs de televisión, la metodología de prueba a utilizar es la de doble estímulo con escala de calidad continua que se describe en el § 3 del anexo 1. La secuencia fuente original se utilizará como condición de referencia. Se sigue debatiendo a propósito de la duración de la secuencia de presentación. En pruebas recientes efectuadas en códecs para vídeo en componentes con relación 4:2:2, se consideró ventajoso modificar la presentación con respecto a la que se da en la presente Recomendación. Se utilizaron imágenes compuestas como referencia adicional para proporcionar un nivel de calidad inferior contra el cual juzgar el comportamiento del códec.

2.2 Evaluación de la calidad en el tratamiento posterior

Con esta evaluación se pretende facilitar la realización de apreciaciones sobre la idoneidad de un códec para aplicaciones de contribución con respecto a un determinado tratamiento posterior, por ejemplo la incrustación cromática, la cámara lenta o el «zoom» electrónico. La disposición de equipo mínima necesaria para tal evaluación consiste en un paso único a través del códec sometido a prueba, seguido del tratamiento posterior objeto de interés y a continuación, el observador. Sin embargo, puede ser más representativo de una aplicación de contribución el empleo de códecs adicionales después del tratamiento posterior.

2.2.1 Condiciones de observación y elección de los observadores

Véase el § 2.1.1.

2.2.2 Utilización de secuencias de imágenes de prueba

Debido a las limitaciones de las posibilidades prácticas de tener que evaluar un códec con varios tratamientos posteriores, el número de secuencias de imágenes de prueba utilizadas puede ser como mínimo de tres, y una más disponible a efectos de demostración, por la imposición de tipo práctico de tener que evaluar probablemente un códec con varios tratamientos posteriores. La naturaleza de las secuencias dependerá de la tarea de tratamiento posterior que se estudie, pero debe variar entre moderadamente crítica y crítica en el contexto de reducción de la velocidad binaria de televisión y para el proceso que se considere. Las secuencias deberán tener una duración del orden de 10 s, pero debe señalarse que los evaluadores pueden preferir una duración de 15 a 30 s. Para la evaluación de la cámara lenta puede servir una velocidad de visualización que sea la décima parte de la de origen.

2.2.3 Metodología de la prueba

La metodología de la prueba que debe utilizarse es la de doble estímulo con escala de calidad continua. Sin embargo aquí la condición de referencia es la fuente sometida al mismo tratamiento posterior que las imágenes decodificadas. Si se considera ventajoso incluir una referencia de calidad inferior también ella deberá someterse al mismo tratamiento posterior. En las pruebas efectuadas por el CCIR se ha introducido una ligera modificación de la presentación que se da en esta Recomendación.

3. Evaluación subjetiva de la degradación de las imágenes de códecs debida a errores de transmisión

Una medida subjetiva útil puede ser la degradación determinada como una función de la velocidad a la que se producen los bits erróneos de transmisión en el enlace entre el codificador y el decodificador. En la actualidad no se tiene conocimiento experimental suficiente de estadísticas ciertas de errores de transmisión, que permitan recomendar parámetros para un modelo que tenga en cuenta las agrupaciones o ráfagas de errores. En tanto no se disponga de esta información, pueden utilizarse los errores con la distribución de Poisson.

3.1 Utilización de secuencias de imágenes de prueba

Limitaciones de tipo práctico inducen a pensar que probablemente serán adecuadas tres secuencias de imágenes de prueba más una de demostración, puesto que hace falta explorar el comportamiento del códec con diversas proporciones de bits erróneos de trasmisión. Cada secuencia debe tener una duración del orden de 10 s, pero debe señalarse que los evaluadores pueden preferir una duración de 15 a 30 s. Estas deben variar entre moderadamente críticas y críticas en el contexto de reducción de la velocidad binaria de televisión.

3.2 Elección de las proporciones de bits erróneos

Deben elegirse al menos cinco proporciones de bits erróneos, pero preferiblemente más, con separación aproximadamente logarítmica y que abarquen la gama que provoca las degradaciones de códec desde «imperceptible» a «muy molesta».

3.3 Metodología de la prueba

Puesto que las pruebas abarcan la gama completa de degradaciones, el método de escala de degradación con doble estímulo es el apropiado y el que debe utilizarse.

3.4 Observación a propósito de la utilización de proporciones de bits erróneos muy bajas

Es posible que haga falta evaluar códecs con proporciones de bits erróneos de transmisión que provoquen transitorias visibles tan infrecuentes que no quepa esperar que se produzcan durante un periodo de secuencias de prueba de 10 s. El tiempo de presentación que aquí se sugiere es claramente inadecuado para tales pruebas.

Si es preciso grabar la salida de un códec en condiciones de proporción de bits erróneos bastante baja (lo que da lugar a un número pequeño de transitorios visibles en un periodo de 10 s) para montaje posterior en presentaciones de evaluación subjetiva, se debe tener la precaución de asegurarse de que la grabación utilizada es típica de la salida del códec observada en un intervalo de tiempo mayor.

4. Comparación subjetiva entre códecs

Cuando no hace falta una apreciación de la calidad o la degradación absolutas de un códec sino sólo su orden de jerarquía, o cuando se desea la confirmación del orden de jerarquía obtenido a partir de los resultados del método de doble estímulo, se debe utilizar el método de las comparaciones de pares de estímulos.

Tal como allí se describe, el método proporciona una comparación sensible y una manera de medir la relación entre pares de sistemas. Es posible una extensión del método, para jerarquizar las calidades o las degradaciones de más de dos sistemas. En este enfoque, el orden de jerarquía global se deriva de la jerarquización de todos los pares posibles de secuencias de imágenes efectuada por los observadores.

El análisis se complica por el hecho de que un observador puede, por ejemplo, clasificar a la imagen A como mejor que la imagen B, y a la imagen B mejor que la C, pero también a la C mejor que la imagen A. Es lo que se denomina una «triada intransitiva».

El número de presentaciones necesarias aumenta con el cuadrado del número de secuencias de imágenes de prueba y de códecs, lo cual representa una desventaja de este método que puede llegar a hacerlo impracticable.

5. Elección del material de imágenes de prueba para la evaluación de los códecs digitales

A lo largo de este anexo se ha hecho énfasis en la importancia de comprobar los códecs digitales con secuencias de imágenes que sean críticas en el contexto de la reducción de la velocidad binaria en televisión. Parece por ello razonable preguntarse en qué medida es crítica una secuencia de imágenes particular para un objetivo determinado de reducción de la velocidad binaria, o si una secuencia es más crítica que otra. Una respuesta sencilla, aunque no especialmente útil, sería decir que «criticidad» significa cosas muy distintas para diferentes códecs. Por ejemplo, podría ocurrir que una imagen fija que contuviera muchos detalles resultase crítica para un códec intratrama mientras que para un códec intercuadro, que es capaz de aprovechar similitudes de cuadro a cuadro, esa misma escena no representaría ninguna dificultad. Algunas secuencias que emplean textura móvil y movimiento complejo resultan críticas para toda clase de códecs, por lo que estos tipos de secuencias son los que más interesa generar o identificar. El movimiento complejo puede tomar la forma de movimientos que son predecibles para un observador pero no para los algoritmos de codificación, como por ejemplo un movimiento periódico tortuoso.

Un examen de posibles medidas estadísticas de criticidad de imagen, por ejemplo mediante métodos correlativos, métodos espectrales, métodos de entropía condicional, etc., ha puesto de manifiesto una medida sencilla pero útil basada en una medición de entropía autoadaptable intratrama/intercuadro. Este método se empleó en la «calibración» de secuencias de imágenes propuestas para utilización en las pruebas de códecs para 34, 45 y 140 Mbit/s en el CCIR y demostró su utilidad para la selección de secuencias empleadas. La manera más sencilla de efectuar tales mediciones en secuencias de imágenes consiste en transferirlas a computadores de procesamiento de imágenes y someterlas a análisis por soporte lógico.

A continuación se dan algunas directrices de carácter general sobre cómo elegir material crítico, para el caso en que no se pueda acceder a las técnicas anteriores.

a) *Códecs intratrama de longitud de palabra fija*

Aunque es posible y válido evaluar estos códecs con imágenes fijas, se recomienda el empleo de secuencias móviles puesto que con ellas resulta más fácil observar los tratamientos del ruido de codificación y son más representativas de las aplicaciones de televisión. Si se emplean imágenes fijas en simulaciones de códecs por computador, se debe efectuar el tratamiento en toda la secuencia de evaluación, para preservar aspectos temporales de cualquier ruido de origen, por ejemplo. Las escenas elegidas deben contener el mayor número posible de los siguientes detalles: zonas estáticas con ciertas texturas y en movimiento (algunas con textura coloreada), objetos estáticos y en movimiento con bordes bruscos de alto contraste de diversas orientaciones (algunos de color); zonas estáticas uniformes semigrises. Del conjunto de secuencias, al menos una debe presentar ruido de origen justamente perceptible y por lo menos una debe ser sintética (es decir, generada por computador) de modo que esté libre de imperfecciones de cámara tales como la abertura de exploración y persistencia de imagen.

b) *Códecs intercuadro de longitud de palabra fija*

Todas las escenas de prueba elegidas deben contener movimiento y el mayor número posible de los siguientes detalles: zonas con ciertas texturas y en movimiento (algunas coloreadas), objetos con bordes bruscos de alto contraste moviéndose en dirección perpendicular a esos bordes y con diversas orientaciones (algunos coloreados). Del conjunto de secuencias, al menos una debe tener ruido de origen justamente perceptible y por lo menos una debe ser sintética.

c) *Códecs intratrama de longitud de palabra variable*

Se recomienda que estos códecs se prueben con material de secuencias de imágenes en movimiento, por las mismas razones que los códecs de longitud de palabra fija. Hay que tener en cuenta que debido a su codificación, de longitud de palabra variable y su memoria intermedia asociada, estos códecs pueden distribuir dinámicamente la capacidad de bits de codificación a través de la imagen. Así por ejemplo, si en la mitad de una imagen se presenta un cielo sin rasgos especiales que no necesita muchos bits para su codificación, se ahorra capacidad para otras partes de la imagen que pueden así reproducirse con calidad elevada, incluso si son críticas. La conclusión importante de todo esto es que si una secuencia de imágenes resulta crítica para un códec de este tipo, habrá que detallar el contenido de cada parte de la pantalla. Debe llenarse con textura en movimiento y estática, con tanta variación de color como se pueda y objetos con bordes bruscos de alto contraste. Al menos una secuencia del conjunto de prueba debe presentar ruido de origen justamente perceptible y por lo menos una debe ser sintética.

d) *Códecs intercuadro de longitud de palabra variable*

Este es el tipo de códec más complejo y el que necesita el material más exigente para forzarlo. No sólo hay que llenar cada parte de la escena con detalles como en el caso del códec intratrama de longitud de palabra variable, sino que esos detalles deben además estar en movimiento. Por otra parte, puesto que muchos códecs emplean métodos de compensación de movimiento, el movimiento a través de la secuencia debe ser complejo. Ejemplos de movimiento complejo son: escenas que emplean simultáneamente el «zoom» y las tomas con movimiento panorámico de la cámara, una escena que tenga como fondo una cortina agitada por el viento y en la que se aprecien sus detalles o su textura; una escena que contenga objetos que giran en un entorno tridimensional; escenas con objetos detallados que se aceleren a través de la pantalla. En todas las escenas debe abundar el movimiento de objetos con diferentes velocidades, texturas y bordes de alto contraste así como un contenido de color variado. Por lo menos una secuencia del conjunto de prueba debe tener ruido de origen justamente perceptible, al menos una debe tener movimiento complejo de cámara generado por computador a partir de una imagen fija natural (de modo que esté libre de ruido y persistencia de imagen de la cámara), y una secuencia cuando menos debe ser generada completamente por computador.

Las secuencias de prueba necesarias para las evaluaciones de tratamiento posterior están sujetas exactamente a los mismos criterios de criticidad. Sin embargo, es posible que resulte difícil cumplir con esos criterios en el caso de secuencias de primeros planos de incrustación cromática porque normalmente tienen una proporción importante de fondo azul sin rasgos característicos.

Se ha preparado una amplia biblioteca de material de secuencias de prueba en formato de componentes con relación 4:2:2, que está grabada en cinta D1. En la Recomendación 802 se dan detalles a propósito de estas secuencias, junto con los criterios con los cuales se prepararon (que pueden aplicarse a otras normas de televisión).