Unión Internacional de Telecomunicaciones

UIT-R

Sector de Radiocomunicaciones de la UIT

Recomendación UIT-R BT.500-13
(01/2012)

Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión

Serie BT

Servicio de radiodifusión (televisión)

Unión Internacional de Telecomunicaciones

Prólogo

El Sector de Radiocomunicaciones tiene como cometido garantizar la utilización racional, equitativa, eficaz y económica del espectro de frecuencias radioeléctricas por todos los servicios de radiocomunicaciones, incluidos los servicios por satélite, y realizar, sin limitación de gamas de frecuencias, estudios que sirvan de base para la adopción de las Recomendaciones UIT-R.

Las Conferencias Mundiales y Regionales de Radiocomunicaciones y las Asambleas de Radiocomunicaciones, con la colaboración de las Comisiones de Estudio, cumplen las funciones reglamentarias y políticas del Sector de Radiocomunicaciones.

Política sobre Derechos de Propiedad Intelectual (IPR)

La política del UIT-R sobre Derechos de Propiedad Intelectual se describe en la Política Común de Patentes UIT-T/UIT-R/ISO/CEI a la que se hace referencia en el Anexo 1 a la Resolución UIT-R 1. Los formularios que deben utilizarse en la declaración sobre patentes y utilización de patentes por los titulares de las mismas figuran en la dirección web http://www.itu.int/ITU-R/go/patents/es, donde también aparecen las Directrices para la implementación de la Política Común de Patentes UIT-T/UIT-R/ISO/CEI y la base de datos sobre información de patentes del UIT-R sobre este asunto.

Series de las Recomendaciones UIT-R

(También disponible en línea en **http://www.itu.int/publ/R-REC/es)**

Series	Título
BO	Distribución por satélite
BR	Registro para producción, archivo y reproducción; películas en televisión
BS	Servicio de radiodifusión sonora
BT	**Servicio de radiodifusión (televisión)**
F	Servicio fijo
M	Servicios móviles, de radiodeterminación, de aficionados y otros servicios por satélite conexos
P	Propagación de las ondas radioeléctricas
RA	Radioastronomía
RS	Sistemas de detección a distancia
S	Servicio fijo por satélite
SA	Aplicaciones espaciales y meteorología
SF	Compartición de frecuencias y coordinación entre los sistemas del servicio fijo por satélite y del servicio fijo
SM	Gestión del espectro
SNG	Periodismo electrónico por satélite
TF	Emisiones de frecuencias patrón y señales horarias
V	Vocabulario y cuestiones afines

***Nota**: Esta Recomendación UIT-R fue aprobada en inglés conforme al procedimiento detallado en la Resolución UIT-R 1.*

Publicación electrónica
Ginebra, 2012

RECOMENDACIÓN UIT-R BT.500-13

Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión

(Cuestión UIT-R 81/6)

(1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002-2009-2012)

Cometido

En la presente Recomendación se describen metodologías para la evaluación de la calidad de la imagen, incluidos métodos generales de prueba, escalas de apreciación y condiciones de observación. Se recomiendan en ella el método de escala de degradación con doble estímulo (DSIS) y el método de escala de calidad continua de doble estímulo (DSCQS), así como otros métodos de evaluación, entre ellos los métodos de estímulo único, los métodos de comparación de estímulos, los métodos de evaluación de calidad continua de estímulo único (SSCQE) y los métodos de doble estímulo simultáneo para evaluación continua (SDSCE).

La Asamblea de Radiocomunicaciones de la UIT,

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 definición mejorada (TVDM) y la TVAD) podrían requerir cambios de los métodos de evaluación subjetiva;

f) que la introducción de dicho procesamiento, señales y servicios aumentará la probabilidad de que la calidad del funcionamiento de cada sección en la cadena de la señal venga condicionada por procesos realizados en partes anteriores de la cadena,

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 en los Anexos 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 § 4 y 5 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;

4 que, para facilitar el intercambio de información entre los distintos laboratorios, los datos recopilados se procesen de acuerdo con las técnicas estadísticas indicadas en el Anexo 2 a la presente Recomendación.

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 una descripción de las técnicas estadísticas empleadas en el procesamiento de los datos recopilados durante las pruebas subjetivas.

Anexo 1

Descripción de los métodos de evaluación

1 Introducción

Se utilizan métodos de evaluación subjetiva para determinar la calidad de funcionamiento de los 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, las 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. Éstas se denominan típicamente evaluaciones de degradación.

Para efectuar evaluaciones subjetivas adecuadas, en primer lugar es necesario seleccionar entre las distintas opciones disponibles aquélla que se adapte mejor a los objetivos y circunstancias del problema de evaluación inmediato. Para ayudar en esta tarea, el § 2 presenta las características generales y en el § 3 aparece información sobre los problemas de evaluación considerados por cada método. A continuación, se detalla en los § 4 y 5 los dos métodos principalmente recomendados. Por último, en el § 6 figura información general sobre métodos alternativos que están siendo sometidos a estudio.

El objeto del presente Anexo se limita a una descripción detallada de los métodos de evaluación. No obstante, la elección del método más adecuado depende de los objetivos del servicio que debe prestar el sistema sometido a prueba. En consecuencia, los procedimientos de evaluación completa de las aplicaciones específicas figuran en otras Recomendaciones UIT-R.

2 Características comunes

Se indican las condiciones generales de observación para las evaluaciones subjetivas. Las condiciones específicas de observación para evaluaciones subjetivas de sistemas concretos figuran en las Recomendaciones conexas.

2.1 Condiciones generales de observación

Se describen distintos entornos con diferentes condiciones de observación.

El entorno de observación de laboratorio tiene por objeto proporcionar condiciones críticas para comprobar el funcionamiento de los sistemas. En el § 2.1.1 se indican las condiciones generales de observación para efectuar evaluaciones subjetivas en el entorno del laboratorio.

El entorno de observación doméstico tiene por objeto proporcionar los medios para evaluar la calidad en el lado de usuario de toda la cadena de transmisión de televisión. Las condiciones generales de observación señaladas en el § 2.1.2 reproducen un entorno próximo al doméstico. Estos parámetros se han seleccionado para definir un entorno ligeramente más crítico que las situaciones normales de observación en los hogares.

Se discuten algunos aspectos relativos a la resolución y el contraste de los monitores.

2.1.1 Entorno de laboratorio

2.1.1.1 Condiciones generales de observación para efectuar evaluaciones subjetivas en el entorno de laboratorio

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

a)	Relación entre la luminancia de pantalla inactiva y el valor de cresta de la luminancia:	≤ 0,02
b)	Relación entre la luminancia de la pantalla, cuando sólo se muestra el nivel del negro en una sala completamente oscura,y la correspondiente al blanco más intenso:	≈ 0,01
c)	Brillo y contraste de la imagen:	Establecido vía PLUGE (véanse las Recomendaciones UIT-R BT.814 y UIT-R BT.815)
d)	Ángulo máximo de observación con respecto al normal (este valor se aplica a las pantallas de tubo de rayos catódicos (TRC), para otro tipo de pantallas se están estudiando los valores adecuados):	30°
e)	Relación entre la luminancia de fondo detrás del receptor de imágenes y el valor de cresta de luminancia de la imagen:	≈ 0,15
f)	Cromaticidad del fondo:	D_{65}
g)	Otra iluminación de la sala:	Débil

2.1.2 Entorno doméstico

2.1.2.1 Condiciones generales de observación para efectuar evaluaciones subjetivas en el entorno doméstico

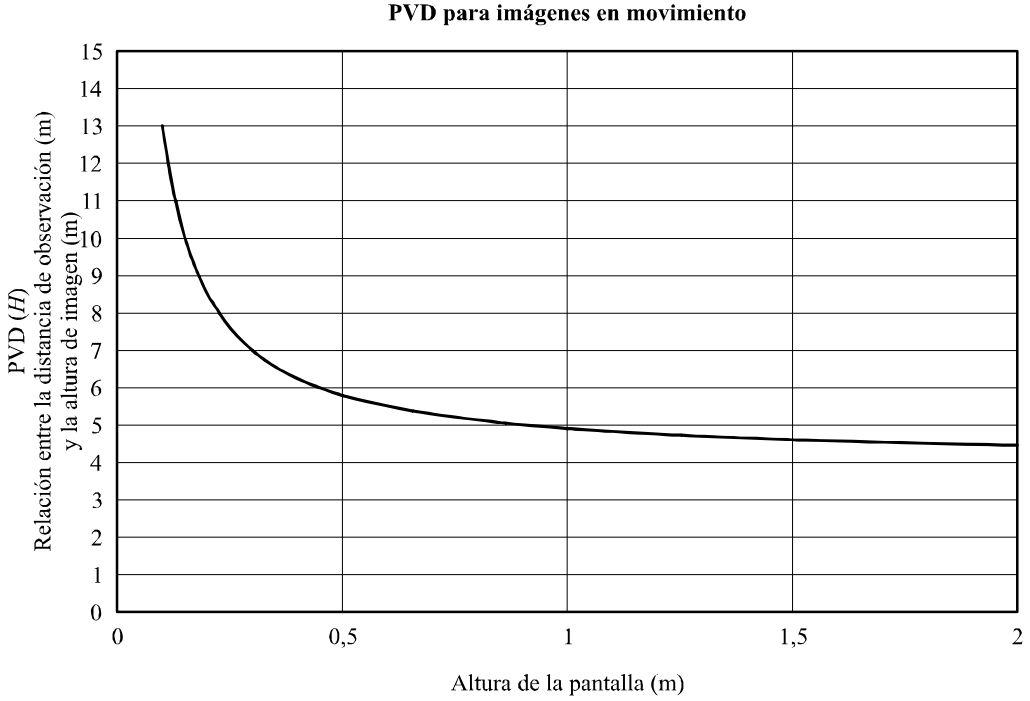
a)	Relación entre la luminancia de pantalla inactiva y el valor de cresta de luminancia:	≤ 0,02 (véase el § 2.1.4)
b)	Brillo y contraste de la imagen:	Establecido vía PLUGE (véanse las Recomendaciones UIT-R BT.818 y UIT-R BT.815)
c)	Ángulo máximo de observación con respecto al normal (este número se aplica a las pantallas de TRC, para otro tipo de pantallas se están estudiando los valores adecuados):	30°

d)	Tamaño de pantalla para un formato de imagen 4/3:	Este tamaño de pantalla debe satisfacer las reglas de la distancia de observación preferida (PVD, *preferred viewing distance)*
e)	Tamaño de pantalla para un formato de imagen 16/9:	Este tamaño de pantalla debe satisfacer las reglas de la PVD
f)	Procesamiento en el monitor:	Sin procesamiento digital
g)	Resolución del monitor:	Véase el § 2.1.3
h)	Valor de cresta de la luminancia:	200 cd/m^2
i)	Luminancia del medio ambiente en la pantalla (Luz incidente del entorno proyectada sobre la pantalla; debe medirse perpendicularmente a la misma):	200 lux

La distancia de observación y los tamaños de pantalla deben elegirse de manera que se satisfaga la PVD. En el Cuadro y el Gráfico siguientes aparecen los valores de la PVD (en función de los tamaños de pantalla). Las cifras son válidas tanto para televisión de definición convencional (TVDC) como para la TVAD, pues las diferencias encontradas son muy pequeñas.

El Cuadro y el Gráfico siguientes ofrecen información sobre la PVD y los tamaños de pantalla correspondientes que deben adoptarse en las Recomendaciones para aplicaciones específicas.

Diagonal de la pantalla (pulgadas)		**Altura de la pantalla (*H*)**	**PVD**
Formato 4/3	**Formato 16/9**	**(m)**	**(*H*)**
12	15	0,18	9
15	18	0,23	8
20	24	0,30	7
29	36	0,45	6
60	73	0,91	5
> 100	> 120	> 1,53	3-4



2.1.3 Resolución del monitor

La resolución de los monitores profesionales, equipados con TRC profesionales, normalmente satisface las normas requeridas para realizar evaluaciones subjetivas en su gama de funcionamiento de luminancia.

No todos los monitores pueden alcanzar un valor de cresta de la luminancia de 200 cd/m^2.

Para verificar y constatar las resoluciones máxima y mínima (en el centro y en las esquinas de la pantalla) puede sugerirse el empleo de un valor de luminancia determinado.

Si para efectuar las evaluaciones subjetivas se utilizan aparatos de televisión domésticos con TRC convencionales la resolución podría ser inadecuada, dependiendo del valor de la luminancia.

En este caso se recomienda encarecidamente verificar y constatar las resoluciones máxima y mínima (en el centro y en las esquinas de la pantalla) para el valor de luminancia utilizado.

Actualmente, la mayoría de los sistemas prácticos disponibles para efectuar evaluaciones subjetivas, a fin de comprobar la resolución de los monitores o de los aparatos de televisión domésticos, utilizan un diagrama de prueba de barrido generado electrónicamente.

Un análisis visual permite verificar la resolución. Se considera que el umbral visual está comprendido entre –12 y –20 dB. El inconveniente principal de este sistema es el repliegue del espectro creado por la máscara que hace difícil la evaluación visual pero, por otro lado, la presencia de repliegue del espectro indica que la señal de videofrecuencia rebasa los límites indicados por la máscara, con muestras de la señal de vídeo.

Es conveniente realizar más estudios sobre pruebas para la definición de los TRC.

2.1.4 Contraste del monitor

El contraste puede venir fuertemente influenciado por la luminancia del entorno.

Los TRC de los monitores profesionales raramente hacen uso de tecnologías para mejorar su contraste en un entorno de alta luminancia, *por lo tanto es posible que no cumplan la norma de contraste necesaria si se utilizan en entornos de alta luminancia*.

Los TRC domésticos emplean tecnologías para conseguir un mejor contraste en un entorno de alta luminancia.

Para calcular el contraste de un TRC determinado, es necesario conocer el coeficiente de reflexión de pantalla, *K*, de dicho tubo. En el mejor caso, el coeficiente de reflexión de pantalla es aproximadamente $K = 6\%$.

Con un entorno difuso de luminancia *I* de 200 lux y un valor de $K = 6\%$, se ha calculado una reflexión de luminancia de 3,82 cd/m^2 en las zonas de pantalla inactivas mediante la siguiente fórmula:

$$L_{reflejada} = \frac{I}{\pi} K$$

Con los valores indicados, la luminancia reflejada (cd/m^2) supone casi el 2% de la luminancia incidente (lux).

Se considera que el TRC no presenta reflexiones especulares en el vidrio frontal, cuya influencia exacta sobre el contraste es difícil de cuantificar porque depende en gran medida de las condiciones de iluminación.

En los § 2.1.1 y 2.1.2 se expresa la relación de contraste, *RC*, de la forma siguiente:

$$RC = L_{mín} / L_{máx}$$

siendo:

$L_{mín}$: luminancia de zonas inactivas en condiciones de iluminación ambiente (cd/m^2) (con los valores indicados: $L_{mín} = L_{zonas\ inactivas} + L_{reflejada} = 3{,}82$ cd/m^2)

$L_{máx}$: luminancia de zonas blancas en condiciones de iluminación ambiente (cd/m^2) (con los valores indicados: $L_{máx} = L_{blanco} + L_{reflejado} = 200 + 3{,}82$ cd/m^2).

Con esos valores se determina una $RC = 0{,}018$, muy próxima al valor de 0,02 indicado en a) de los § 2.1.1.1 y 2.1.2.1.

2.2 Señales fuente

La señal fuente 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 fuente más reproducibles, y son por consiguiente las preferidas. Pueden intercambiarse entre laboratorios, para dar mayor significado a las comparaciones de sistemas. Se pueden utilizar formatos de cintas de computador o vídeo.

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 convenientes. 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.

Frecuentemente será necesario tener en cuenta la forma en que pueden afectar a la calidad de funcionamiento del sistema sometido a prueba los efectos de cualquier procesamiento realizado en una etapa anterior de la señal. En consecuencia, es conveniente que siempre que se lleven a cabo pruebas en secciones de la cadena que puedan dar lugar a distorsiones de procesamiento, aunque no sean visibles, la señal resultante debe ser grabada de forma transparente y a continuación debe dejarse disponible para pruebas posteriores, cuando se desea determinar cómo pueden acumularse a lo largo de la cadena las degradaciones debidas a un procesamiento en cascada. Dichas grabaciones deben almacenarse en la biblioteca del material de prueba, para futura utilización si es preciso, y deben incluir una indicación detallada de los precedentes de la señal grabada.

2.3 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 1 se describen los problemas de evaluación y de materiales de prueba típicos utilizados para abordar esos problemas.

CUADRO 1

Selección del material de prueba*

Problema de evaluación	Material utilizado
Calidad de funcionamiento global con material de uso habitual	General, «crítico pero no en exceso»
Capacidad, aplicaciones críticas (por ejemplo, contribución, postprocesamiento, etc.)	Diverso, incluido el material muy crítico para la aplicación probada
Calidad de funcionamiento de sistemas «adaptables»	Material muy crítico para el esquema «adaptable» utilizado
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 al Anexo 1 se pueden obtener mayores directrices para la selección de materiales de prueba.

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 UIT-R. En la Recomendación relativa a la evaluación de las aplicaciones se propone material de imágenes específico.

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

2.4 Gama de condiciones y anclaje

Dado que la mayoría de los métodos de evaluación 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).

2.5 Observadores

Los observadores pueden ser expertos o no expertos dependiendo de los objetivos de la evaluación. Un observador experto cuenta con experiencia en las perturbaciones de la imagen que puede introducir el sistema sometido a prueba. Un observador no experto no tiene esta experiencia. En todo caso, los observadores no deben estar directamente familiarizados con el sistema sometido a prueba; es decir, no deben tener conocimientos específicos y detallados sobre el mismo.

Antes de una sesión, debe examinarse a los observadores para determinar su agudeza visual normal (o corregida) mediante los gráficos de Snellen o Landolt y su visión normal de los colores, utilizando gráficos elegidos especialmente (por ejemplo, los de Ishihara). El número de asesores necesarios depende de la sensibilidad y la fiabilidad del procedimiento de prueba adoptado y del tamaño previsto del efecto que se busca. Para estudios de alcance limitado, por ejemplo, de carácter exploratorio, pueden emplearse menos de 15 observadores. En tales casos, el estudio debe considerarse «informal» y debe comunicarse el nivel de experiencia de los observadores en la evaluación de la calidad de la imagen de televisión.

Según un estudio de la coherencia entre los resultados de los diferentes laboratorios de prueba, se pueden producir diferencias sistemáticas entre los resultados obtenidos por los distintos laboratorios. Tales diferencias serán particularmente importantes si se pretende agregar los resultados de diversos laboratorios para mejorar la sensibilidad y la fiabilidad de un experimento.

La explicación de las diferencias entre los diversos laboratorios podría hallarse quizás en los distintos niveles de destreza de los diferentes grupos de evaluadores. Es preciso seguir investigando para saber hasta qué punto es cierta esta hipótesis y, si se demuestra que lo es, cuantificar las variaciones imputables a ese factor. Mientras tanto, los experimentadores deberán incluir el mayor número de detalles posible sobre las características de sus equipos de evaluación, para facilitar la investigación a propósito del referido factor. Entre los datos que podrían proporcionarse figuran los de categoría laboral (por ejemplo, empleado de organización radiodifusora, estudiante de universidad, empleado de oficina, ...), sexo y edad.

2.6 Instrucciones para la evaluación

Debe familiarizarse detenidamente a los evaluadores con el método de evaluación, el factor de calidad, los tipos de degradaciones que probablemente se produzcan, la escala de apreciaciones, la secuencia y la temporización. Las secuencias de entrenamiento que demuestran la gama y el tipo de degradaciones que van a evaluarse deben emplearse con imágenes ilustrativas distintas a las

utilizadas en las pruebas, pero de sensibilidad comparable. En el caso de evaluaciones de la calidad, puede definirse ésta como un conjunto de atributos perceptuales específicos.

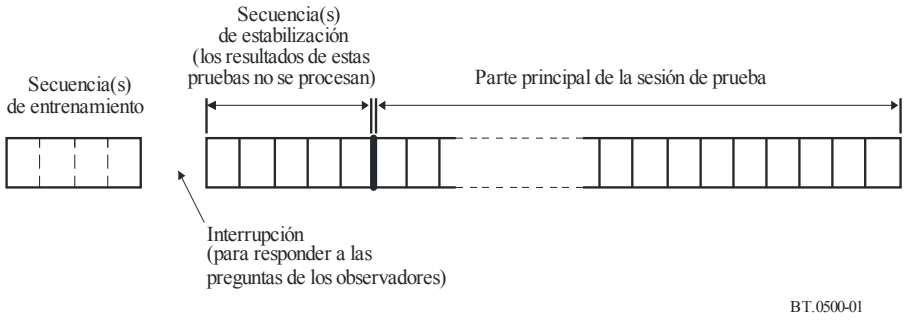
2.7 Sesión de prueba

Una sesión debe durar al menos media hora. Al principio de la primera sesión, deben realizarse unas cinco «presentaciones fingidas» para estabilizar la opinión de los observadores. Los datos obtenidos de estas presentaciones no deben tenerse en cuenta en los resultados de la prueba. Si se necesitan varias sesiones, sólo es preciso realizar tres presentaciones fingidas al principio de la siguiente sesión.

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.

FIGURA 1

Estructura de presentación de la sesión de prueba



2.8 Presentación de los resultados

Como varían con la gama, es inadecuado interpretar las apreciaciones a partir de la mayoría de los métodos de evaluación en términos absolutos (por ejemplo, la calidad de una imagen o secuencia de imágenes).

Para cada parámetro de prueba debe darse la media y el intervalo de confianza del 95% 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. En el Anexo 2 a la presente Recomendación aparece más información sobre procesamiento de datos.

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 fuente y de los monitores (véase la Nota 1),
– número y tipo de evaluadores (véase la Nota 2),
– sistemas de referencias utilizados,
– nota media global del experimento,
– notas media original y ajustada, e intervalo de confianza del 95% si se ha eliminado uno o más observadores de acuerdo con un procedimiento.

NOTA 1 – Puesto que existe cierta evidencia en el sentido de que el tamaño de la pantalla puede influir en los resultados de los evaluadores subjetivos, se pide a los experimentadores que notifiquen de manera explícita las dimensiones de la pantalla, así como la marca y el número de modelo de los dispositivos de presentación visual utilizados en cualquier experimento.

NOTA 2 – Se ha comprobado que las variaciones en el grado de destreza de los equipos de observadores (incluso entre equipos de «no especializados» pueden influir en los resultados de las evaluaciones de observación subjetivas. Para facilitar un ulterior estudio de este factor, se pide a los experimentadores que comuniquen el mayor número posible de las características de sus equipos de observación. Podrían ser factores de interés los siguientes: la composición, en cuanto a edad y sexo, del equipo o bien su nivel educativo o categoría laboral.

3 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 2 se describen los problemas de evaluación característicos y los métodos utilizados para abordar dichos problemas.

CUADRO 2

Selección del método de prueba

Problema de evaluación	Método utilizado	Descripción
Medir la calidad de los sistemas con respecto a una referencia	Método de escala de calidad continua de doble estímulo (DSCQS)[1]	Rec. UIT-R BT.500, § 5
Medir la robustez de los sistemas (es decir, características de fallo)	Método de escala de degradación con doble estímulo (DSIS)[1]	Rec. UIT-R BT.500, § 4
Cuantificar la calidad de los sistemas (cuando no se dispone de referencias)	Método de valoración cuantitativa[2] o valoración categórica (en estudio)	Informe UIT-R BT.1082
Comparar la calidad de sistemas alternativos (cuando no se dispone de referencias)	Método de comparación directa método de valoración cuantitativa[2] o valoración categórica (en estudio)	Informe UIT-R BT.1082
Identificar factores en los que se observa que los sistemas difieren y medir su influencia perceptual	En estudio	Informe UIT-R BT.1082
Establecer el punto en el cual una degradación se hace visible	Estimación del umbral por el método de elección forzada o método de ajuste (en estudio)	Informe UIT-R BT.1082
Determinar si se perciben diferencias en los sistemas	Método de elección forzada (en estudio)	Informe UIT-R BT.1082
Medir la calidad de la codificación de imagen estereoscópica	Método de escala de calidad continua de doble estímulo (DSCQS)[3]	Rec. UIT-R BT.500, § 5
Medir la fidelidad entre dos secuencias vídeo degradadas	Método de doble estímulo simultáneo para evaluación continua (SDSCE)	Rec. UIT-R BT.500, § 6.4
Comparar diferentes instrumentos de elasticidad a errores	Método de doble estímulo simultáneo para evaluación continua (SDSCE)	Rec. UIT-R BT.500, § 6.4

(1) Se han llevado a cabo algunos estudios sobre efectos contextuales para el método DSCQS y el método DSIS. Se ha determinado que los resultados del método DSIS presentan tendencias sistemáticas en un cierto grado debido a los efectos contextuales. En el Apéndice 3 al Anexo 1, aparecen mayores detalles al respecto.

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

(3) Debido a la posibilidad de que aparezca una intensa fatiga cuando se evalúan imágenes estereoscópicas, la duración total de la sesión de evaluación debe ser inferior a 30 min.

4 Método de escala de degradación con doble estímulo (DSIS) (método UER)

4.1 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 § 5).

El método de 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.

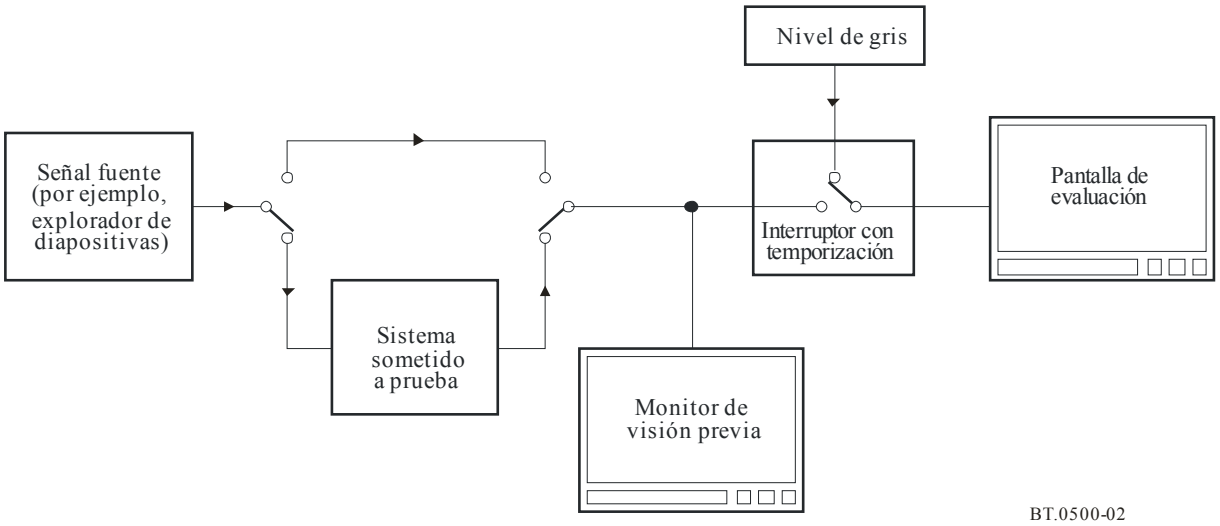
4.2 Disposición general

En el § 2 se indica la forma de definir o seleccionar las condiciones de observación, las señales fuente, el material de prueba y los observadores así como la presentación de los resultados.

La disposición general del sistema de prueba debería ser la que se indica en la Fig. 2.

FIGURA 2

Disposición general de los sistemas de prueba para el método de DSIS



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 fuente, 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 de la fuente, y la segunda es la misma imagen encaminada por el sistema sometido a prueba.

4.3 Presentación del material de prueba

Una sesión de prueba consta de varias presentaciones. Hay dos variantes de la estructura de las presentaciones, la I y la II que se indican a continuación:

Variante I: La imagen o secuencia de referencia y la imagen o secuencia de prueba se presentan sólo una vez, como muestra la Fig. 3a).

Variante II: La imagen o secuencia de referencia y la imagen o secuencia de prueba se presentan dos veces, como muestra la Fig. 3b).

La variante II, que tiene una mayor duración que la variante I, puede aplicarse si es necesario discriminar entre degradaciones muy pequeñas o se están sometiendo a prueba secuencias en movimiento.

4.4 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.

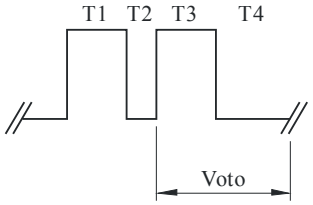
4.5 Introducción a las evaluaciones

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.

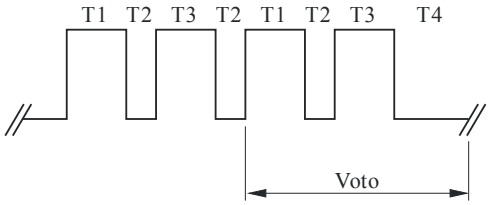
Debe pedirse a los observadores que observen la imagen durante los periodos T1 y T3. La votación debe autorizarse únicamente durante T4.

FIGURA 3

Estructura de presentación del material de prueba



a) Variante I



b) Variante II

Fases de presentación:

T1 =	10 s	Imagen de referencia
T2 =	3 s	Gris mediano producido por un nivel vídeo de unos 200 mV
T3 =	10 s	Condición a evaluar
T4 =	5-11 s	Gris mediano

La experiencia sugiere que prolongar los periodos T1 y T3 más alla de 10 S no mejora la capacidad del evaluador para juzgar las imágenes o las secuencias.

BT.0500-03

4.6 La sesión de prueba

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 al Anexo 1 se presentan otras ideas sobre la selección de niveles de degradaciones.

5 El método de escala de calidad continua de doble estímulo (DSCQS)

5.1 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.

5.2 Disposición general

En el § 2 se indica la forma de definir o seleccionar las condiciones de observación, las señales fuente, el material de prueba, los observadores y la introducción a la evaluación. La sesión de prueba se describe en el § 4.6.

La disposición general del sistema de prueba debería ser la que se indica en la Fig. 4.

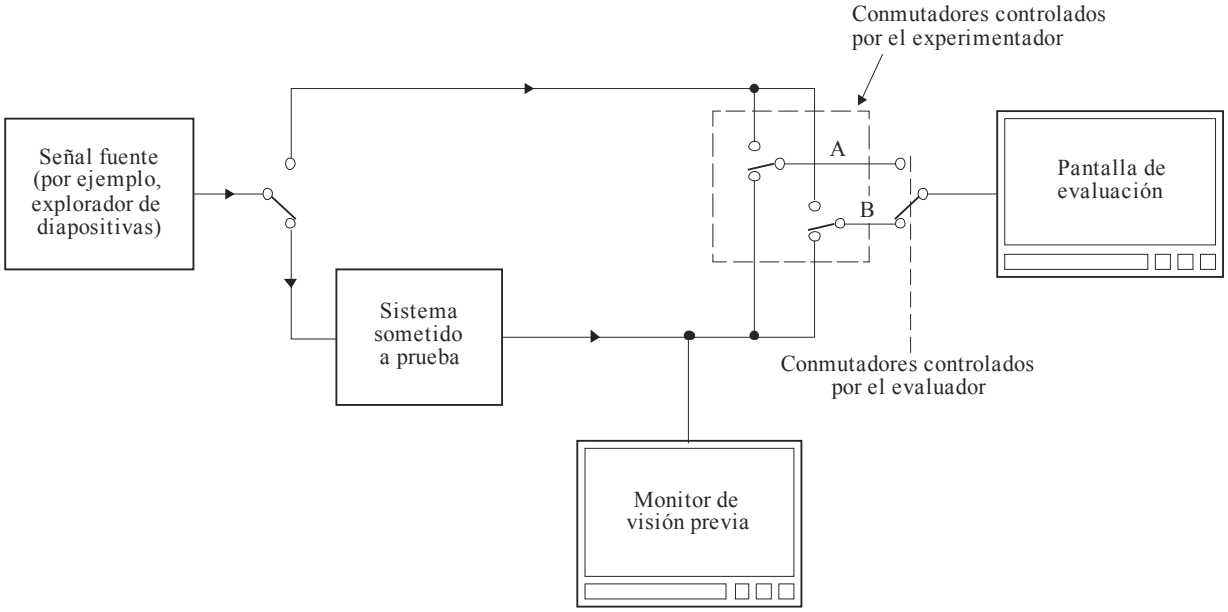
5.3 Presentación del material de prueba

Una sesión de prueba 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). La estructura de las presentaciones se muestra en la Fig. 5.

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 fuente invertida en el tiempo a través del sistema que se está probando.

FIGURA 4

Disposición general del sistema de prueba para el método DSCQS



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

Variante I: El evaluador, que suele estar solo, puede conmutar enre 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.

Variante 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.

BT.0500-04

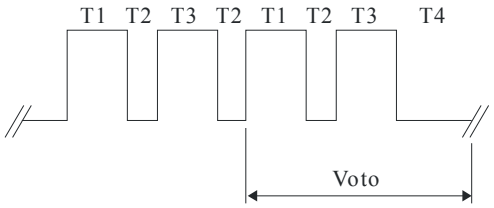
5.4 Escala de apreciación

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 UIT-R. 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. 6 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.

FIGURA 5

Estructura de presentación del material de prueba



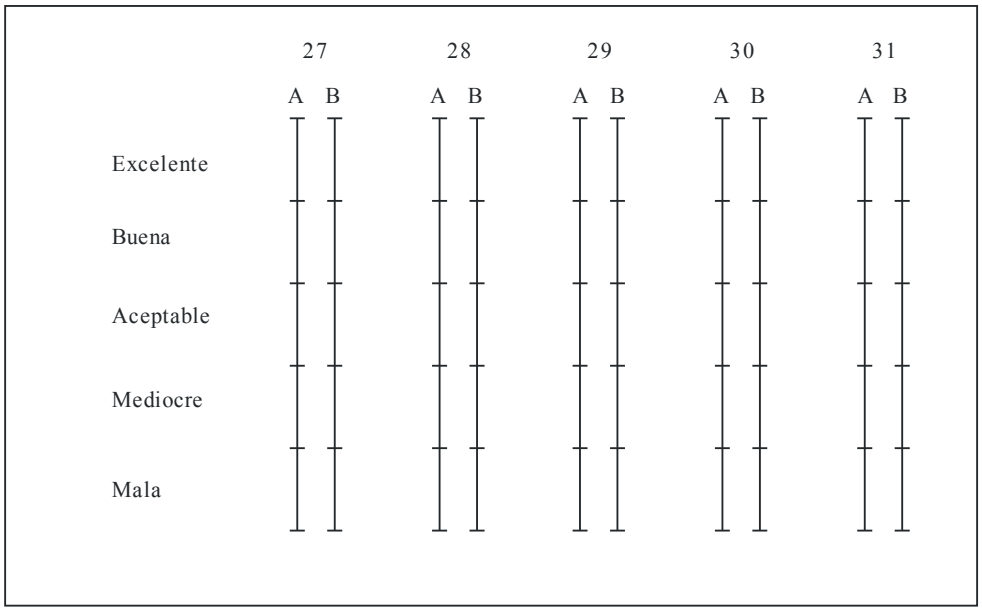
Fases de presentación:

T1 = 10 s Secuencia de prueba A
T2 = 3 s Gris mediano producido por un nivel vídeo de unos 200 mV
T3 = 10 s Secuencia de prueba B
T4 = 5-11 s Gris mediano

BT.0500-05

FIGURA 6

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



* Al planificar la disposición de los elementos de prueba en una sesión de evaluación para el método DSCQS conviene que el experimentador incluya verificaciones para asegurar que el experimento carece de errores sistemáticos. Sin embargo, el método para llevar a cabo estas verificaciones aún es objeto de investigación.

BT.0500-06

5.5 Análisis de los resultados

Los pares de evaluaciones (de referencia y de prueba) correspondientes a 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. A continuación se calculan las diferencias entre la evaluación de la condición de referencia y la de prueba. En el Anexo 2 se describen otros procedimientos.

La experiencia ha mostrado que los resultados obtenidos para diferentes secuencias de prueba dependen de la criticidad del material de prueba utilizado. Se puede conseguir una interpretación más completa de la calidad de funcionamiento del códec presentando los resultados de diferentes

secuencias de prueba de manera separada, en vez de presentarlos simplemente como medias acumuladas de todas las secuencias de prueba utilizadas en la evaluación.

Si los resultados de las secuencias de prueba se disponen en una clasificación por categoría de «criticidad de la secuencia de prueba» en un eje de abscisas, es posible presentar una descripción gráfica aproximada de la característica de fallo de la imagen según el contenido del sistema sometido a prueba. Sin embargo, esta forma de presentación sólo describe la calidad de funcionamiento del códec, no proporciona ninguna indicación de la probabilidad de que se produzcan secuencias con un grado determinado de criticidad (véase el Apéndice 1 al Anexo 1). Es preciso seguir estudiando la criticidad de las secuencias de prueba y la probabilidad de que se produzcan secuencias con un determinado nivel de criticidad antes de que se pueda obtener esta imagen más completa del funcionamiento del sistema.

5.6 Interpretación de los resultados

Cuando se utiliza este método DSCQS, podría ser arriesgado e incluso erróneo deducir conclusiones a propósito de la calidad de las condiciones de prueba asociando valores de DSCQS numéricos a adjetivos procedentes de otros protocolos de prueba (por ejemplo, imperceptible, perceptible, pero no molesta, ... tomados del método DSIS).

Se señala que los resultados obtenidos por el método DSCQS no deberán tratarse como resultados absolutos sino como diferencias de resultados entre una condición de referencia y una condición de prueba. Así pues, es erróneo asociar los resultados a un solo término de descripción de calidad, incluso con los que proceden del propio protocolo DSCQS (por ejemplo, excelente, buena, aceptable, ...).

En cualquier procedimiento de prueba es importante establecer criterios de aceptabilidad antes de comenzar la evaluación. Esto tiene una importancia especial cuando se utiliza el método de DSCQS debido a la tendencia de los usuarios poco expertos a interpretar erróneamente el significado de los valores de la escala de calidades producidos por el método.

6 Otros métodos de evaluación

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

6.1 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. El material de prueba podría consistir únicamente en secuencias de prueba o en secuencias de prueba con sus correspondientes secuencias de referencia. En este último caso, la secuencia de referencia se presenta como estímulo independiente para generar índices como cualquier otro estímulo de prueba.

6.1.1 Disposición general

En el § 2 se indica la forma de definir o seleccionar las condiciones de observación, las señales fuente, la gama de condiciones y anclaje, los observadores, la introducción a la evaluación y la presentación de los resultados.

6.1.2 Selección del material de prueba

Para las pruebas de laboratorio debe seleccionarse el contenido de las imágenes de prueba como se describe en el § 2.3.

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).

6.1.3 Sesión de prueba

La sesión de prueba consiste en una serie de pruebas de evaluación, que deberían presentarse según un orden aleatorio y, preferiblemente, en una secuencia aleatoria distinta para cada observador. Cuando se utiliza un orden aleatorio único de secuencias, hay dos variantes de la estructura de las presentaciones: I (estímulo único) y II (estímulo único con repetición múltiple) como se indica a continuación:

a) Las imágenes o secuencias de prueba se presentan solamente una vez en la sesión de prueba; al comienzo de las primeras sesiones deberán introducirse algunas secuencias fingidas (descritas en el § 2.7). El experimentador se asegura normalmente de que la misma imagen se presente dos veces seguidas con el mismo nivel de degradación.

   Una prueba de evaluación típica consiste en tres presentaciones: un campo de adaptación en gris medio, un estímulo y un campo de post-exposición en gris medio. Las duraciones de esas presentaciones varían según la tarea del observador, los materiales y las opiniones 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.

b) Las imágenes o secuencias de prueba se presentan tres veces organizando la sesión de prueba en tres presentaciones, cada una de las cuales incluye todas las imágenes de secuencias que se han de probar solamente una vez; el comienzo de cada presentación se anuncia mediante un mensaje en el monitor (por ejemplo, Presentación 1). La primera presentación se utiliza para estabilizar la opinión del observador; los datos generados por esta presentación no se deben tener en cuenta en los resultados de la prueba; las notas asignadas a las imágenes o secuencias se obtienen promediando los datos generados por las presentaciones segunda y tercera. El experimentador se asegura normalmente de que se aplican las siguientes limitaciones al orden aleatorio de las imágenes o secuencias dentro de cada presentación:

   – una determinada imagen o secuencia no está en la misma posición en las demás presentaciones;

   – una determinada imagen o secuencia no está situada inmediatamente antes de la misma imagen o secuencia en las demás presentaciones.

Una prueba de evaluación típica consiste en dos presentaciones: un estímulo y un campo de post-exposición en gris medio. Las duraciones de esas presentaciones pueden variar según la tarea del observador, los materiales y las opiniones o factores examinados, no obstante se sugieren duraciones de 10 y 5 s respectivamente. El índice o los índices del observador pueden recogerse durante la presentación del campo de post-exposición únicamente.

La variante II (estímulo único con repetición múltiple) introduce claramente una tara en el tiempo requerido para efectuar una sesión de prueba (45 s frente a 23 s para cada imagen o secuencia que se prueba); no obstante, disminuye la fuerte dependencia de los resultados de la variante I con respecto al orden de las imágenes o secuencias dentro de una sesión.

Además, los resultados de los experimentos muestran que la variante II permite un margen de fluctuación en torno al 20% dentro de la gama de los votos.

6.1.4 Tipos de métodos de estímulo único

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

6.1.4.1 Métodos de apreciación por categorías de adjetivos

En las apreciaciones por categorías de adjetivos, 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 UIT-R se dan en el Cuadro 3. En controles operacionales se utilizan a veces medias notas. Las escalas que evalúan la legibilidad del texto, el esfuerzo de lectura, y la utilidad de la imagen se han utilizado en casos especiales.

CUADRO 3

Escalas de calidad y degradación del UIT-R

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

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.

6.1.4.2 Métodos de apreciación por categorías numéricas

Se ha estudiado un procedimiento de estímulo único que utiliza una escala de categoría numérica de once notas (SSNCS) y se ha comparado con las escalas gráficas y cuantitativas. Este estudio, descrito en el Informe UIT-R BT.1082, señala una clara preferencia por el método SSNCS, en términos de sensibilidad y estabilidad, cuando no se dispone de referencia.

6.1.4.3 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 3). 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 numérica, 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).

6.1.4.4 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.

6.2 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.

6.2.1 Disposición general

En el § 2 se indica la forma de definir o seleccionar las condiciones de observación, las señales de origen, la gama de condiciones y anclaje, los observadores, la introducción a la evaluación y la presentación de los resultados.

6.2.2 Selección del material 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.

6.2.3 Sesión de prueba

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 las 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.

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.

6.2.4 Tipos de métodos de comparación de estímulos

En las evaluaciones de televisión se han utilizado los tres tipos de métodos de comparación de estímulos.

6.2.4.1 Métodos de apreciación por categorías de adjetivos

En los métodos de apreciación por categorías de adjetivos, 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 UIT-R se indica en el Cuadro 4.

CUADRO 4

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.).

6.2.4.2 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 4). 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.

6.2.4.3 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.

6.3 Evaluación de calidad continua de estímulo único (SSCQE)

La introducción de la compresión en la televisión digital provocará degradaciones de la calidad de la imagen dependientes de la escena y variables con el tiempo. Incluso dentro de breves muestras de vídeo codificado digitalmente, la calidad puede variar mucho dependiendo del contenido de la escena y las degradaciones pueden ser de muy corta duración. Las metodologías convencionales del UIT-R no bastan por sí solas para evaluar este tipo de material. Además, el método del doble estímulo de prueba de laboratorio no reproduce las condiciones de observación doméstica de estímulo único. Por ello, se ha considerado conveniente que la calidad subjetiva del vídeo codificado digitalmente se mida de manera continua, observando los sujetos participantes el material una sola vez, sin una referencia fuente.

Como resultado de lo anterior, se ha elaborado y probado la siguiente técnica nueva SSCQE.

6.3.1 Evaluación continua de la calidad global

6.3.1.1 Dispositivo de registro y configuración

Se ha de utilizar un sistema de registro electrónico conectado a un computador para registrar la evaluación de calidad continua por parte de los participantes. Este dispositivo deberá tener las características siguientes:

– su mecanismo deslizante no ha de tener ninguna posición armada,
– la distancia de desplazamiento lineal ha de ser de 10 cm,
– fijo o montado en consola,
– las muestras se han de registrar dos veces por segundo.

6.3.1.2 Formato general del protocolo de prueba

A los participantes se les presentarán sesiones de prueba con el siguiente formato:

– *Segmento de programa*: un segmento de programa corresponde a un tipo de programa (por ejemplo, deportes, noticias, teatro) procesado de acuerdo con uno de los parámetros de calidad objeto de evaluación (por ejemplo, la velocidad binaria); cada segmento de programa debe durar por lo menos 5 min;
– *Sesión de prueba*: una sesión de prueba es una serie de una o más combinaciones diferentes de segmento de programa/parámetro de calidad sin separación y dispuestas en orden seudoaleatorio. Cada sesión de prueba contiene por lo menos una vez todos los segmentos de programa y parámetros de calidad, pero no necesariamente todas las combinaciones segmento de programa/parámetro de calidad; cada sesión de prueba deberá durar entre 30 y 60 min;
– *Presentación de prueba*: una presentación de prueba representa la realización completa de una prueba. Se puede dividir una presentación de prueba en sesión de prueba para cumplir con los requisitos de duración máxima y para evaluar la calidad con todos los pares de segmentos de programa/parámetros de calidad. Si el número de pares segmento de programa/parámetro de calidad es limitado, se puede hacer una presentación de prueba repitiendo la misma sesión de prueba, para que la prueba dure un periodo de tiempo suficientemente largo.

Se puede introducir audio a efectos de evaluación de la calidad del servicio. En este caso, la selección del material audio de acompañamiento deberá efectuarse atribuyéndole la misma importancia que a la selección del material vídeo, antes de realizar la prueba.

En el formato de prueba más sencillo se utilizaría un solo segmento de programa y se tendría en cuenta un solo parámetro de calidad.

6.3.1.3 Parámetros de observación

Las condiciones de observación deberán ser las especificadas actualmente en las Recomendaciones UIT-R BT.500, UIT-R BT.1128, UIT-R BT.1129 y UIT-R BT.710.

6.3.1.4 Escalas de apreciación

Al dar las instrucciones de la prueba a los participantes, deberá quedar claro que la distancia de desplazamiento del mecanismo deslizante del microteléfono corresponde a la escala de calidad continua descrita en el § 5.4.

6.3.1.5 Observadores

Deberán participar al menos 15 observadores, no especializados, con las características que actualmente se recomiendan en el § 2.5.

6.3.1.6 Instrucciones a los observadores

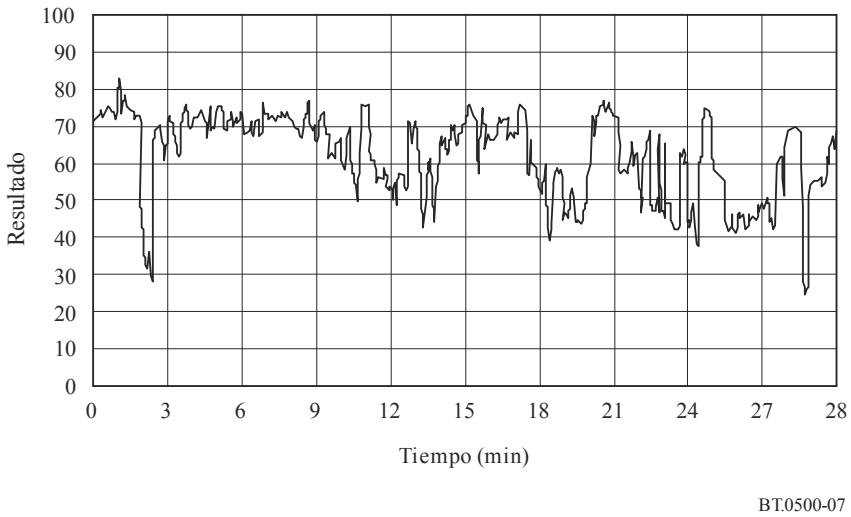
Si se evalúa la calidad de servicio (con audio de acompañamiento), deberá indicarse a los observadores que tengan en cuenta la calidad global, en vez de fijarse en la calidad vídeo solamente.

6.3.1.7 Presentación de datos y procesamiento y presentación de resultados

Deberán recogerse datos de todas las sesiones de prueba. De esta manera será posible obtener un gráfico único del índice de calidad media en función del tiempo, $q(t)$, como media de las apreciaciones de la calidad de todos los observadores por segmento de programa, parámetro de calidad o sesión de prueba completa (véase el ejemplo de la Fig. 7).

FIGURA 7

Condición de prueba: códex X/segmento de programa Z



BT.0500-07

Sin embargo, la variabilidad del tiempo de respuesta de los diferentes observadores puede influir en los resultados de la estimación si el promedio se calcula solamente en un segmento de programa. Se están llevando a cabo estudios para evaluar la influencia del tiempo de respuesta de los diferentes observadores en la apreciación de calidad resultante.

Los datos anteriores pueden convertirse a un histograma de probabilidad de la ocurrencia del nivel de calidad q, $P(q)$ (véase el ejemplo de la Fig. 8).

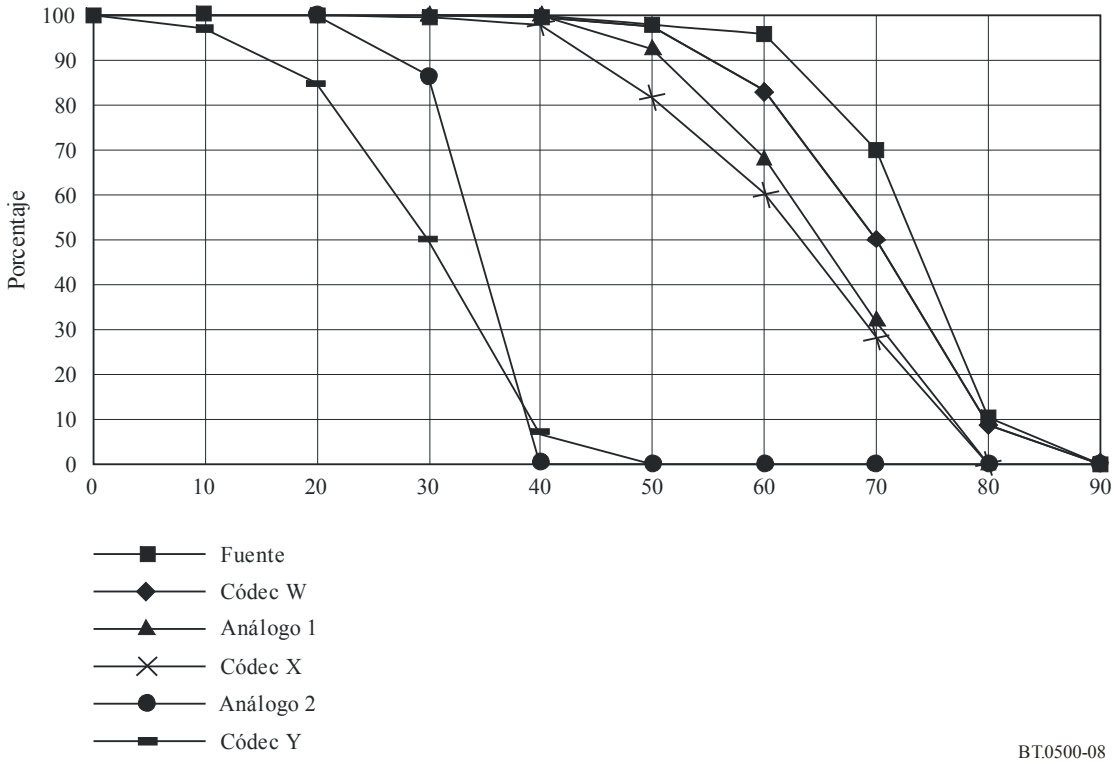
6.3.2 Calibración de los resultados de calidad continuos y obtención de un único índice de calidad

Aunque existen pruebas de que pueden producirse sesgos basados en la memoria, en sesiones largas de evaluación de un único índice de calidad de vídeo codificado digitalmente por el método DSCQS, recientemente se ha comprobado que tal efecto no es significativo si las evaluaciones DSCQS se efectúan con muestras de vídeo de 10 s. En consecuencia, una posible segunda etapa del proceso SSCQE, actualmente en estudio, consistiría en calibrar el histograma de calidad utilizando el método DSCQS existente en muestras de 10 s representativas, extraídas de los datos del histograma.

Las metodologías convencionales del UIT-R, empleadas en el pasado, han servido para generar índices de calidad únicos de secuencias de televisión. Se han llevado a cabo experimentos en los que se ha examinado la relación entre la evaluación continua de una secuencia de vídeo codificada y un índice de calidad global único del mismo segmento. Ya se ha visto que los efectos de la memoria humana pueden distorsionar los índices de calidad si se producen degradaciones notables en aproximadamente los últimos 10 a 15 s de la secuencia. Sin embargo, también se ha visto que dichos efectos podrían modelarse como una función de ponderación exponencial descendente. De aquí la posibilidad de una tercera etapa en la metodología SSCQE, que consistiría en procesar los resultados de esas evaluaciones de calidad continuas para obtener una medición de calidad única equivalente. Se trata de algo que está siendo objeto de estudio actualmente.

FIGURA 8

Media de los resultados de las secuencias de votación sobre el segmento de programa Z



6.4 Método de doble estímulo simultáneo para evaluación continua (SDSCE)

La idea de una evaluación continua surgió en el UIT-R porque los métodos anteriores presentaban algunas deficiencias para la medición de la calidad del vídeo de esquemas de compresión digital. Las principales desventajas de los métodos normalizados anteriores están vinculadas a la ocurrencia de artefactos relacionados con el contexto en las imágenes digitales visualizadas. En los protocolos anteriores, la duración del tiempo de observación de las secuencias vídeo en evaluación está limitado generalmente a 10 s, lo que obviamente no es suficiente para que el observador tenga un juicio representativo de lo que pudo suceder en el servicio real. Los artefactos digitales dependen en gran medida del contenido espacial y temporal de la imagen fuente. Esto es válido para los esquemas de compresión pero también en relación con el comportamiento de la elasticidad a los errores de los sistemas de transmisión digital. Con los anteriores métodos normalizados era muy difícil elegir secuencias vídeo representativas, o por lo menos evaluar su representatividad. Por este motivo, el UIT-R introdujo el método SSCQE, que es capaz de medir la calidad vídeo en secuencias más largas, representativas del contenido vídeo y de la estadística de errores. Para reproducir las condiciones de observación que estén lo más próximas posibles a las situaciones reales, en el SSCQE no se utilizan referencias.

Cuando hay que evaluar la fidelidad, se han de introducir condiciones de referencia. El SDSCE ha sido elaborado a partir del SSCQE, con ligeras diferencias en cuanto a la manera de presentar las imágenes a los sujetos y con respecto a la escala de apreciación. El método fue propuesto a MPEG para evaluar la solidez contra los errores a velocidades binarias muy bajas, pero puede ser aplicado adecuadamente a todos los casos en los que hay que evaluar la fidelidad de la información visual afectada por la degradación que varía en función del tiempo.

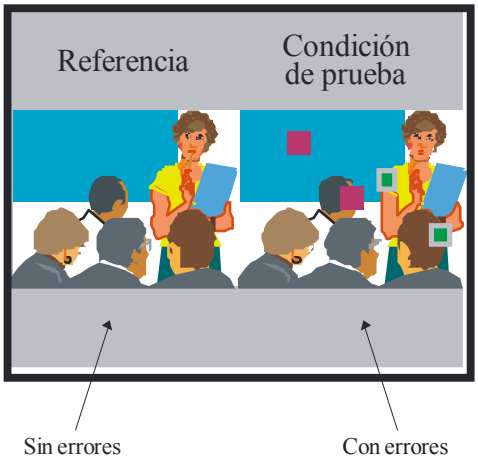
Como resultado, se ha elaborado y probado la siguiente nueva técnica SDSCE.

6.4.1 Procedimiento de prueba

El grupo de sujetos observa dos secuencias al mismo tiempo: una es la referencia, la otra es la condición de prueba. Si el formato de las secuencias es de formato de imagen normalizado (SIF) o más pequeño, las dos secuencias pueden ser visualizadas juntas en el mismo monitor; en los demás casos se debe utilizar dos monitores alineados (véase la Fig. 9).

FIGURA 9

Ejemplo de formato de visualización



BT.0500-09

Se pide a los sujetos que comprueben las diferencias entre las dos secuencias y juzguen la fidelidad de la información vídeo moviendo el cursor de un dispositivo de voto manual. Cuando la fidelidad es perfecta, el cursor debe estar en la parte superior de la escala (codificada 100), cuando la fidelidad es nula, el cursor debe estar en la parte inferior de la escala (codificada 0).

Los sujetos conocen cuál es la referencia y se les pide que expongan su opinión, durante todo el tiempo que están observando las secuencias.

6.4.2 Diferentes fases

La *fase de entrenamiento* es una parte esencial de este método de prueba, porque los sujetos podrían comprender mal su tarea. Se deben proporcionar instrucciones escritas para estar seguros de que todos los sujetos reciben exactamente la misma información. Las instrucciones deben incluir la explicación sobre lo que los sujetos van a ver, lo que tienen que evaluar (es decir, la diferencia de calidad) y cómo tienen que exponer su opinión. Todas las preguntas de los sujetos deben ser respondidas para evitar en la mayor medida posible todo prejuicio de opinión del administrador de la prueba.

Después de las instrucciones, se debe efectuar una *sesión de demostración*. De esta manera los sujetos se familiarizan con los procedimientos de voto y la clase de degradaciones.

Por último, se debe efectuar una prueba simulada, en la cual se muestran varias condiciones representativas. Las secuencias deben ser diferentes de las utilizadas en la prueba y deben ser presentadas una después de otra sin interrupción.

Cuando termina la *prueba simulada*, el experimentador debe comprobar principalmente que en caso de que las condiciones de prueba sean iguales a las referencias, las evaluaciones estén próximas al ciento (es decir, no se ha visto diferencia); si en cambio los sujetos declaran ver algunas diferencias, el experimentador debe repetir la explicación y la prueba simulada.

6.4.3 Características del protocolo de prueba

Las siguientes definiciones se aplican a la descripción del protocolo de prueba:

– *Segmento vídeo*: un segmento vídeo corresponde a una secuencia vídeo.

– *Condición de prueba*: una condición de prueba puede ser un proceso vídeo específico, una condición de transmisión, o ambos. Cada segmento vídeo debe ser procesado de acuerdo con una condición de prueba por lo menos. Además, se deben añadir referencias a la lista de condición de prueba, con el fin de hacer pares de referencia/referencia que se han de evaluar.

– *Sesión*: una sesión es una serie de diferentes segmentos vídeo/condiciones de prueba pares sin separación y arregladas en un orden pseudoaleatorio. Cada sesión contiene por lo menos una vez todos los segmentos vídeo y condiciones de prueba pero no necesariamente todas las combinaciones de segmento vídeo/condición de prueba.

– *Presentación de prueba*: una presentación de prueba es una serie de sesiones para abarcar todas las combinaciones de segmento vídeo/condición de prueba. Todas las combinaciones de segmento vídeo/condición de prueba deben ser votadas por el mismo número de observadores (pero no necesariamente los mismos observadores).

– *Periodo de votación*: se pide a cada observador que vote continuamente durante una sesión.

– *Segmento de votos*: un segmento de 10 s de votos; todos los segmentos de votos se obtienen utilizando grupos de 20 votos consecutivos (equivalentes a 10 s) sin ninguna superposición.

6.4.4 Procesamiento de datos

Una vez efectuada la prueba, uno (o más) ficheros de datos están disponibles con todos los votos de las diferentes sesiones (S) que representan todo el material de voto de la presentación de prueba (TP). Se puede efectuar una primera comprobación de la validez de los datos verificando que cada par de segmentos vídeo/condiciones de prueba ha sido presentado y que un número equivalente de votos ha sido asignado a cada uno de ellos.

Los datos recopilados durante la ejecución de las pruebas realizadas de acuerdo con este protocolo pueden ser procesados de tres maneras diferentes:

– Análisis estadístico de cada segmento vídeo separado.

– Análisis estadístico de cada condición de prueba separada.

– Análisis estadístico global de todos los segmentos vídeo/condiciones de prueba pares.

En cada caso se requiere un análisis de múltiples pasos:

– Se calculan los valores medios y las desviaciones típicas para cada voto por acumulación de los observadores.

– Se calcula el promedio y la desviación típica para cada segmento de votos, como se ilustra en la Fig. 10. Los resultados de este paso pueden ser representados en un diagrama temporal, como se muestra en la Fig. 11.

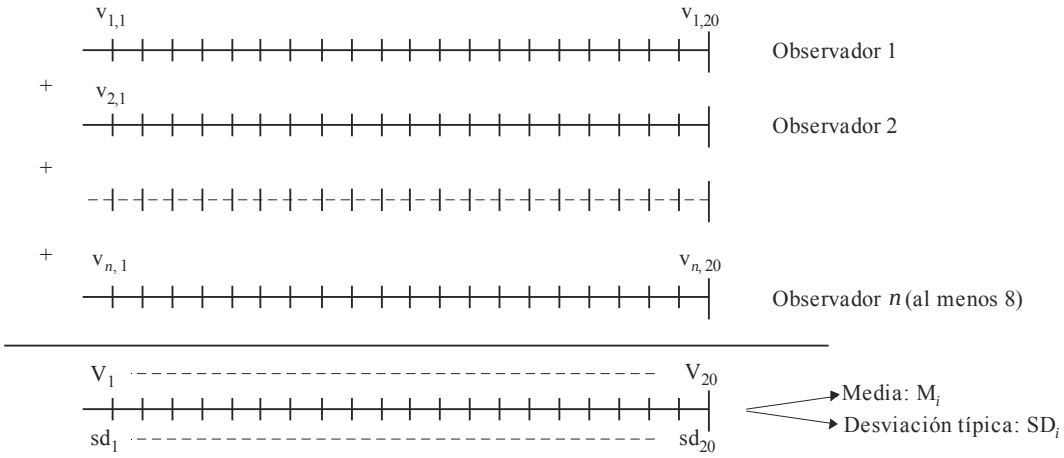
– Se analiza la distribución estadística de los valores medios calculados en el paso anterior (es decir, correspondiente a cada segmento de votos), y su frecuencia de aparición. Para evitar el efecto de novedad debido a las anteriores combinaciones de segmentos vídeo × condiciones de prueba, se rechazan los primeros 10 s de votos para cada muestra de segmento vídeo × condición de prueba.

– La característica global de molestia se calcula acumulando las frecuencias de ocurrencia. En este cálculo se deben tener en cuenta los intervalos de confianza, como se muestra en la Fig. 12. Una característica global de molestia corresponde a esta función de distribución estadística acumulada mostrando la relación entre los valores medios para cada segmento de votación y su frecuencia de aparición acumulada.

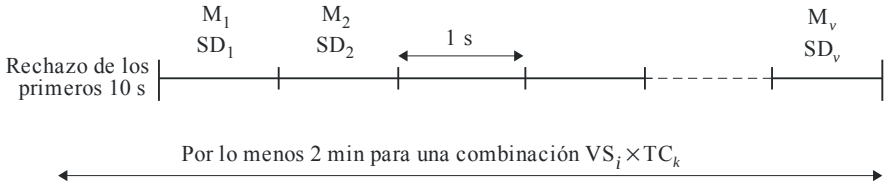
FIGURA 10

Procesamiento de datos

a) Cálculo de la nota media, V, y la desviación típica, SD, por instante de voto de los observadorespara cada secuencia de votación de cada combinación segmento de vídeo (VS) × condición de prueba (TC)



b) Cálculo de M y SD por secuencia de votación de 1 s para cada combinación VS × TC



BT.0500-10

6.4.5 Fiabilidad de los sujetos

La fiabilidad de los sujetos puede ser evaluada cualitativamente comprobando su comportamiento cuando se muestran los pares de referencia/referencia. En estos casos, se espera que los sujetos den evaluaciones muy próximas a 100. Esto prueba que por lo menos han comprendido su tarea y que sus votos no son aleatorios.

Además, la fiabilidad de los sujetos puede ser comprobada utilizando procedimientos que están próximos al descrito en el § 2.3.2 del Anexo 2 para el método SSCQE.

En el procedimiento SDSCE, la fiabilidad de los votos depende de los dos parámetros siguientes:

Desviación sistemática: durante una prueba, un observador puede ser demasiado optimista o demasiado pesimista, o puede incluso haber entendido mal los procedimientos de votación (por ejemplo, el significado de la escala de votación). Esto puede conducir a una serie de votos con desviación sistemática con respecto a la serie media, si no completamente fuera de gama.

Inversiones locales: como en otros procedimientos de prueba muy conocidos, algunas veces los observadores votan sin preocuparse mucho de observar y seguir cuidadosamente la calidad de la secuencia visualizada. En este caso, la curva global de voto puede estar relativamente dentro de la gama media. Sin embargo, es posible observar las inversiones locales.

Estos dos efectos indeseables (comportamiento atípico e inversiones) podrían evitarse. Naturalmente, el entrenamiento de los participantes es muy importante, pero debe ser posible utilizar un instrumento que permita detectar y, si es necesario, descartar a los observadores incoherentes. En esta Recomendación se describe una propuesta de un proceso de dos pasos que permite efectuar este filtrado.

FIGURA 11

Diagrama temporal

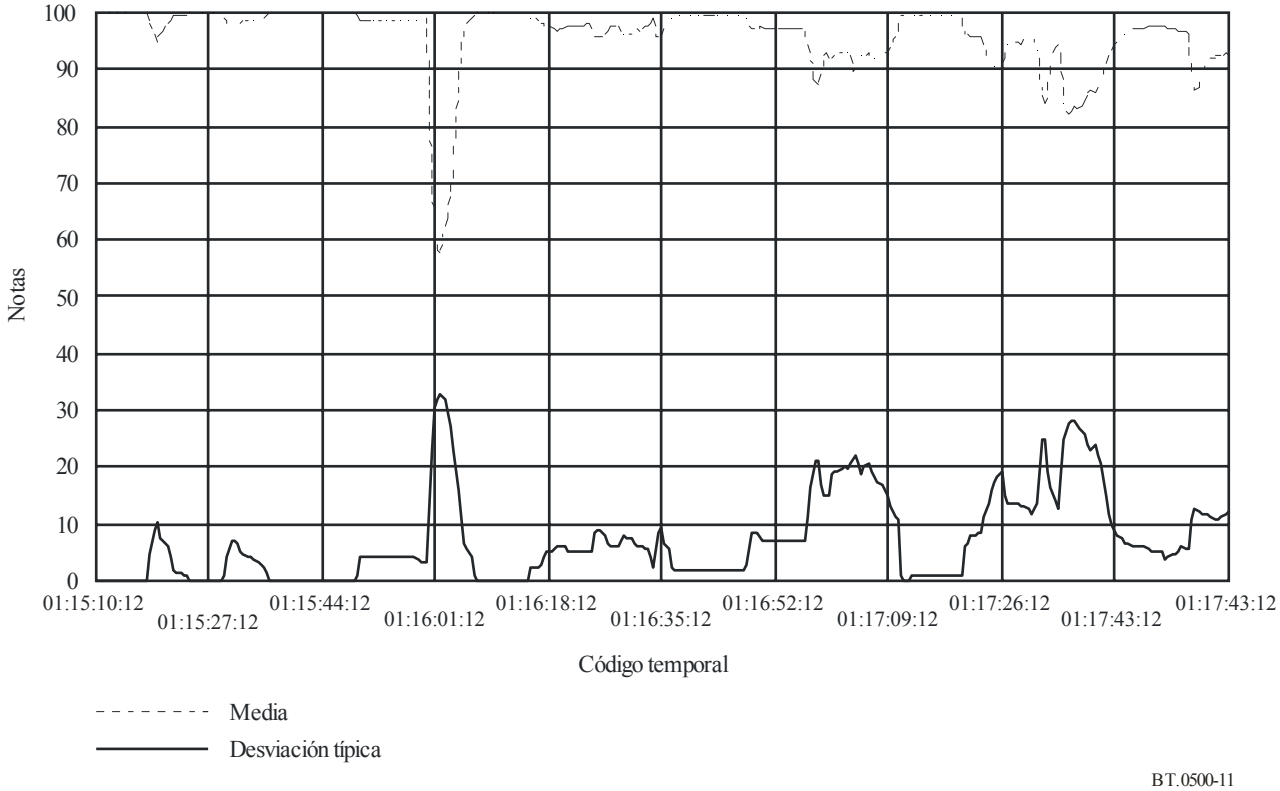
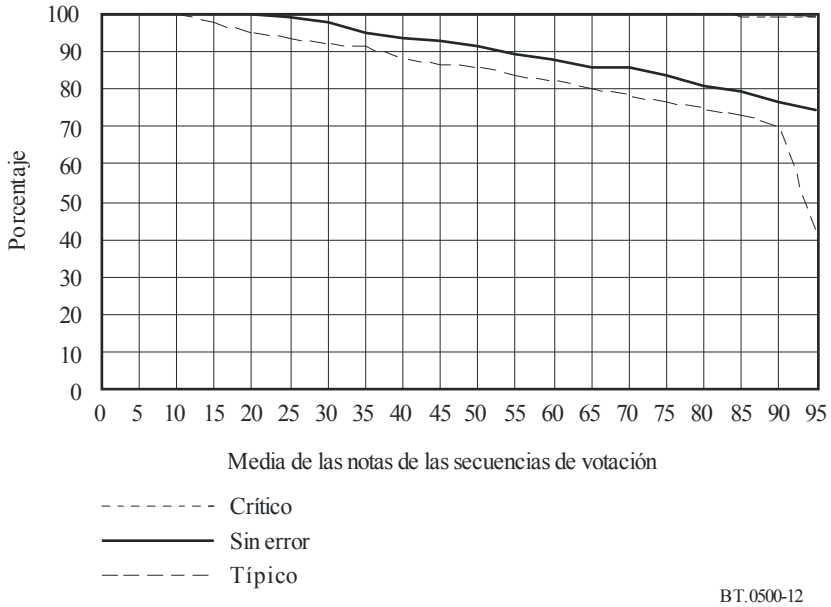


FIGURA 12

Características globales de molestia calculadas a partir de las distribuciones estadísticas e incluido el intervalo de confianza



6.5 Observaciones

En el Informe UIT-R BT.1082 se describen otras técnicas, tales como los métodos con escalas multidimensionales y los métodos de variables múltiples, que aún son objeto de estudio.

Todos los métodos descritos hasta ahora tienen sus ventajas y sus limitaciones, y todavía no es posible recomendar uno preferentemente con carácter definitivo. 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

Característica de fallo de la imagen según su contenido

1 Introducción

Tras su implantación, 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 de la imagen según su contenido 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 la imagen según su contenido 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. 13.

Una característica de fallo de la imagen según su contenido 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. 14 se ilustra un ejemplo de dicha distribución.

– *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. 15 se ilustra un ejemplo de dicha función.

– *Paso 4*: conlleva la información de los Pasos 2 y 3 a fin de obtener una característica de fallo de la imagen según su contenido de la forma indicada en la Fig. 13.

FIGURA 13

Representación gráfica de una característica posible de fallo de la imagen según su contenido

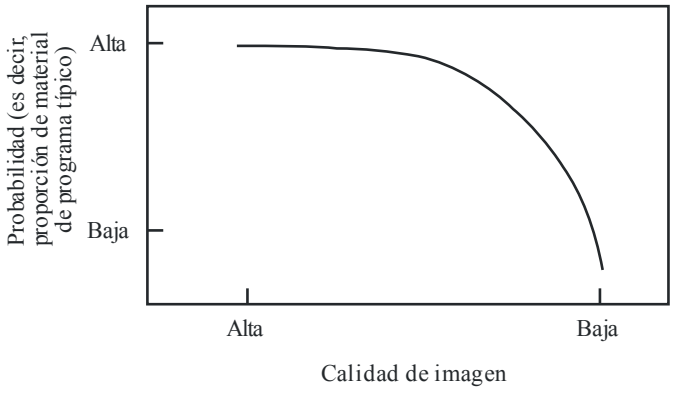


FIGURA 14

Probabilidad de aparición de material de programa con niveles de criticidad diferentes

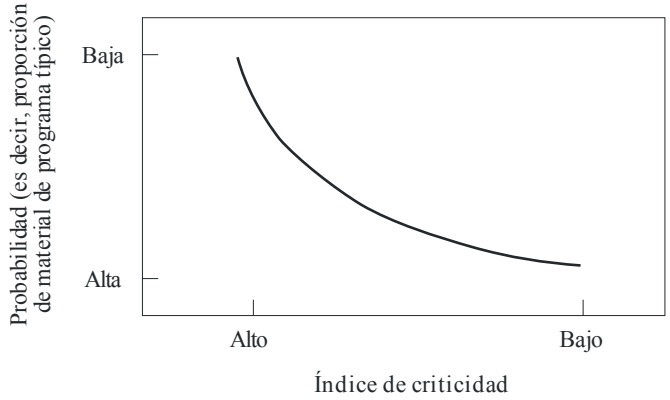
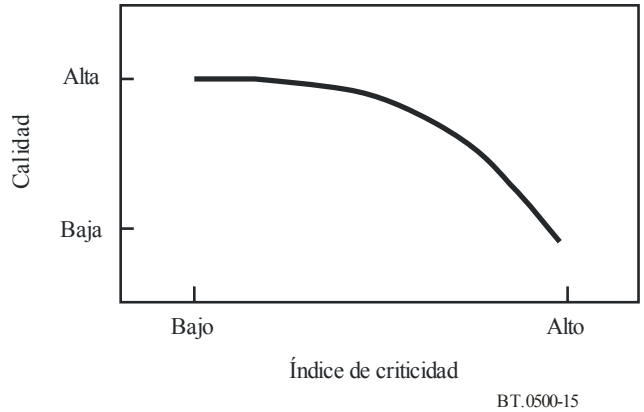


FIGURA 15

Función que relaciona la calidad con la criticidad del material de programa



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 la imagen según su contenido 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.

En el Anexo 1 a la Recomendación UIT-R BT.1129 se da un ejemplo de característica de fallo de la imagen según su contenido para televisión digital.

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;
– 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 200 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 al Anexo 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 § 5 del Anexo 1). 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.

Apéndice 3
al Anexo 1

Efecto contextual

Los efectos contextuales aparecen cuando la calificación subjetiva de una imagen viene influenciada por el orden y la severidad de las degradaciones presentes. Por ejemplo, si se presenta una imagen muy degradada después de un conjunto de imágenes ligeramente degradadas, los observadores pueden calificar inadvertidamente esta imagen con una nota más baja de lo que lo harían normalmente.

Un grupo de cuatro laboratorios de distintos países han investigado los posibles efectos contextuales asociados a los resultados de tres métodos (método DSCQS, método DSIS, variante II y un método de comparación) utilizados para evaluar la calidad de imagen. El material de prueba se obtuvo mediante codificación MPEG (ML@MP) junto con reducción de la resolución horizontal. A cada serie de pruebas, una de ellas sobre degradaciones contextuales débiles y la otra sobre degradaciones intensas, se le aplicaron cuatro condiciones de prueba básicas (B1, B2, B3, B4) y seis condiciones de prueba contextuales. Se aplicaron los tres métodos de prueba a ambas series de pruebas. Los efectos contextuales son la diferencia entre los resultados de la prueba con degradaciones predominantemente débiles y la prueba con fundamentalmente degradaciones predominantemente intensas. Las condiciones de prueba básicas B2 y B3 se utilizaron para determinar los efectos contextuales.

Los resultados combinados de los laboratorios indican que no hay efectos contextuales para el método DSCQS. Para los métodos DSIS y de comparación los efectos contextuales fueron evidentes y el efecto más intenso apareció para el método DSIS, variante II. Los resultados indican que las degradaciones predominantemente débiles pueden provocar calificaciones más bajas de una imagen y las degradaciones predominantemente fuertes pueden provocar calificaciones más elevadas.

Los resultados de la investigación sugieren que el método DSCQS es el más adecuado para minimizar los efectos contextuales en la evaluación subjetiva de la calidad de imagen recomendada por el UIT-R.

En el Informe UIT-R BT.1082 aparece más información sobre este tema.

Anexo 2

Análisis y presentación de los resultados

1 Introducción

En el transcurso de un experimento subjetivo para evaluar la calidad de funcionamiento de un sistema de televisión, se recopila un gran volumen de datos. Estos datos, en forma de hojas de evaluación de los observadores, o su equivalente electrónico, deben condensarse mediante técnicas estadísticas para ofrecer resultados de manera gráfica y/o numérica/ formulada/algorítmica en los que se resume la calidad de funcionamiento del sistema sometido a prueba.

El siguiente análisis es aplicable a los resultados de los métodos de un solo estímulo del método DSIS y del método DSCQS para la evaluación de la calidad de imágenes de televisión (véanse los § 4, 5 y 6 del Anexo 1) y a otros métodos alternativos que utilizan escalas numéricas. En el primer y segundo caso, se evalúa la degradación en una escala de cinco notas o multinota. En el último caso,

se utilizan escalas de evaluación continua y los resultados (diferencias entre la evaluación de la imagen de referencia y la imagen real sometida a prueba) se normalizan a valores enteros comprendidos entre 0 y 100.

2 Métodos comunes de análisis

Las pruebas realizadas de acuerdo con los principios de los métodos descritos en el Anexo 1 producirán una distribución de valores enteros comprendidos, por ejemplo, entre 1 y 5 o entre 0 y 100. Habrá variaciones en estas distribuciones debido a las diferencias de apreciación entre observadores y al efecto de diversas condiciones asociadas al experimento, por ejemplo, la utilización de varias imágenes o de secuencias.

Una prueba constará de varias presentaciones, *L*. Cada presentación de prueba será una de entre varias condiciones de prueba, *J*, aplicada a una de entre varias secuencias de prueba/imágenes de prueba, *K*. En algunos casos, podrá repetirse un cierto número de veces, *R*, cada una de las combinaciones de secuencia de prueba/imagen de prueba y condición de prueba.

2.1 Cálculo de notas medias

El primer paso para analizar los resultados consiste en calcular la nota media, $\bar{u}_{jkr}$, correspondiente a cada una de las presentaciones:

$$\bar{u}_{jkr} = \frac{1}{N} \sum_{i=1}^{N} u_{ijkr} \tag{1}$$

donde:

u_{ijkr}: nota del observador *i* para la condición de prueba *j*, secuencia/imagen *k*, repetición *r*

N: número de observadores.

De manera similar, podrían calcularse las notas medias globales, $\bar{u}_j$ y $\bar{u}_k$, correspondientes a cada condición de prueba y secuencia/imagen de prueba.

2.2 Cálculo del intervalo de confianza

2.2.1 Procesamiento de datos brutos (no compensados y/o no aproximados)

Cuando se presenten los resultados de una prueba, todas las notas medias deberán tener un intervalo de confianza asociado que se obtiene a partir de la desviación típica y el tamaño de cada muestra.

Se propone utilizar un intervalo de confianza del 95%, que viene dado por:

$$\left[\bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr}\right]$$

donde:

$$\delta_{jkr} = 1{,}96 \frac{S_{jkr}}{\sqrt{N}} \tag{2}$$

La desviación típica de cada presentación, S_{jkr}, viene dada por:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{3}$$

Con una probabilidad del 95%, el valor absoluto de la diferencia entre la nota media experimental y la nota media «verdadera» (para un número de observadores muy elevado) es menor que el intervalo de confianza del 95%, siempre que la distribución de las notas individuales cumpla ciertos requisitos.

De manera similar, podría calcularse la desviación típica, S_j, correspondiente a cada condición de prueba. Se señala no obstante que, cuando se utilice un número muy reducido de secuencias de prueba/imágenes de prueba, esta desviación típica se verá influida más por las diferencias entre las secuencias de prueba empleadas que por las variaciones entre los observadores participantes en la evaluación.

2.2.2 Procesamiento de datos compensados y/o aproximados

Para los datos cuyos efectos de degradación/mejora y efectos frontera residuales de la escala de evaluación hayan sido compensados, o los datos presentados en forma de ley de respuesta o adición de degradaciones después de la aproximación, debido a la dependencia de las notas medias experimentales de calidad con respecto a estas distorsiones, el intervalo de confianza deberá calcularse utilizando transformaciones de variables estadísticas teniendo en cuenta la dispersión de la variable correspondiente.

Si los resultados de la evaluación se presentan a modo de respuesta de degradaciones (es decir, como una curva experimental), los límites inferior y superior del intervalo de confianza serán función de los valores experimentales. Para calcular esos límites de confianza se ha de calcular la desviación típica y se ha de evaluar una aproximación de su dependencia para cada valor experimental de la respuesta de degradaciones original.

2.3 Selección de los observadores

2.3.1 Selección para los métodos DSIS, DSCQS y alternativos, salvo el método SSCQE

En primer lugar, se debe examinar si la distribución de las notas para cada presentación 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. Para cada presentación, las notas u_{ijkr} de cada observador deben compararse con el valor medio asociado, $\overline{u}_{jkr}$, más dos veces la desviación típica asociada, S_{jkr} (si es normal) o $\sqrt{20}$ veces (si no es normal) P_{jkr}, y el valor medio asociado menos dos veces la misma desviación típica o $\sqrt{20}$ veces Q_{jkr}. Cada vez que una nota del observador sea superior a P_{jkr} se incrementa un contador asociado a cada observador, P_i. De manera similar, cada vez que una nota del observador sea inferior a Q_{jkr}, se incrementa un contador asociado a cada observador, Q_i. Por último, se deben calcular las dos relaciones siguientes: $P_i + Q_i$ dividido por el número total de notas de cada observador durante la sesión entera, y $P_i - Q_i$ dividido por $P_i + Q_i$ como valor absoluto. Si la primera relación es mayor del 5% y la segunda relación es menor del 30%, se debe rechazar al observador i (véase la Nota 1).

NOTA 1 – Este procedimiento no debe aplicarse más de una vez a los resultados de un experimento determinado. Además, el empleo del procedimiento ha de estar limitado a los casos en los que haya relativamente pocos observadores (por ejemplo, menos de 20), todos ellos no especializados.

Este procedimiento es el que se recomienda para el método UER (DSIS); también se ha aplicado con éxito al método DSCQS y a métodos alternativos.

El proceso anterior puede expresarse matemáticamente de la forma siguiente:

Para cada presentación de prueba, se calcula la media, $\overline{u}_{jkr}$, la desviación típica, S_{jkr}, y el coeficiente de curtosis, β_{2jkr}. Este coeficiente viene dado por:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{con} \quad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \overline{u}_{ijkr})^x}{N} \tag{4}$$

Para cada observador, i, se obtiene P_i y Q_i, es decir:

Para j, k, r = 1, 1, 1 a J, K, R

Si $2 \le \beta_{2jkr} \le 4$, entonces:

si $u_{ijkr} \ge \overline{u}_{jkr} + 2\, S_{jkr}$ entonces $P_i = P_i + 1$

si $u_{ijkr} \le \overline{u}_{jkr} - 2\, S_{jkr}$ entonces $Q_i = Q_i + 1$

o bien:

si $u_{ijkr} \ge \overline{u}_{jkr} + \sqrt{20}\, S_{jkr}$ entonces $P_i = P_i + 1$

si $u_{ijkr} \le \overline{u}_{jkr} - \sqrt{20}\, S_{jkr}$ entonces $Q_i = Q_i + 1$

Si $\dfrac{P_i + Q_i}{J \cdot K \cdot R} > 0{,}05$ y $\left|\dfrac{P_i - Q_i}{P_i + Q_i}\right| < 0{,}3$ se rechaza al observador i

siendo:

N: número de observadores

J: número de condiciones de prueba incluida la de referencia

K: número de imágenes o secuencias de prueba

R: número de repeticiones

L: número de presentaciones de prueba (en la mayoría de los casos, el número de presentaciones será igual a $J \cdot K \cdot R$; no obstante, se señala que algunas evaluaciones pueden llevarse a cabo con números distintos de secuencias para cada condición de prueba).

2.3.2 Selección para el método SSCQE

Para la selección de observadores específicos cuando se utiliza el procedimiento de prueba SSCQE, el dominio de aplicación ya no es una de las configuraciones de prueba (combinación de una condición de prueba y una secuencia de prueba) sino una ventana de tiempo (por ejemplo, un segmento de voto de 10 s) de una configuración de prueba. Se efectúa un filtrado de los participantes en dos pasos, el primero se emplea para detectar y descartar observadores que presenten una discrepancia muy acusada en sus votos en comparación con el comportamiento medio y el segundo se realiza para detectar y seleccionar observadores incoherentes, sin consideración alguna a la discrepancia sistemática en las apreciaciones.

Paso 1: Detección de las inversiones de voto local

En este caso también debe examinarse en primer lugar si la distribución de notas para cada ventana de tiempo de cada configuración de prueba es «normal» o no utilizando la prueba β_2. Si β_2 se encuentra entre 2 y 4, la distribución puede considerarse «normal». En ese caso se aplica el proceso para cada ventana de prueba de cada configuración de prueba como se expresa matemáticamente a continuación.

Para cada ventana de tiempo de cada una de las configuraciones de prueba y utilizando los votos u_{ijkr} de cada observador, se calcula la media $\bar{u}_{jklr}$, la desviación típica, S_{jklr} y el coeficiente, β_{2jklr}. Este coeficiente viene dado por la expresión:

$$\beta_{2jklr} = \frac{m_4}{(m_2)^2} \quad \text{con} \quad m_x = \frac{\sum_{n=1}^{N} (u_{njklr} - \bar{u})^x}{N}$$

Para cada observador, i, se determinan P_i y Q_i, es decir:

Para $j, k, l, r = 1, 1, 1, 1$, a J, K, L, R

Si $2 \le \beta_{2jklr} \le 4$, entonces:

si $u_{njklr} \ge \bar{u}_{jklr} + 2\, S_{jklr}$ entonces $P_i = P_i + 1$

si $u_{njklr} \le \bar{u}_{jklr} + 2\, S_{jklr}$ entonces $Q_i = Q_i + 1$

o bien:

si $u_{njklr} \ge \bar{u}_{jklr} + \sqrt{20}\, S_{jklr}$ entonces $P_i = P_i + 1$

si $u_{njklr} \le \bar{u}_{jklr} - \sqrt{20}\, S_{jklr}$ entonces $Q_i = Q_i + 1$

Si $\frac{P_i}{J \cdot K \cdot L \cdot R} > X\%$ o $\frac{Q_i}{J \cdot K \cdot L \cdot R} > X\%$ se rechaza al observador i

siendo:

N: número de observadores

J: número de ventanas de tiempo en una combinación de prueba de condición y secuencias de prueba

K: número de condiciones de prueba

L: número de secuencias

R: número de repeticiones.

Este proceso permite eliminar observadores que han emitido votos muy distantes de las notas medias. En la Fig. 17 aparecen dos ejemplos (las dos curvas de los extremos presentan discrepancias importantes). No obstante, este criterio de eliminación no permite detectar posibles inversiones que es otra fuente importante de deformaciones sistemáticas en las apreciaciones. Por esa razón se propone un segundo paso.

Paso 2: Detección de inversiones del voto local

En este Paso 2 la detección también se basa en las fórmulas de selección indicadas en el Anexo 2 a la presente Recomendación. Se introduce una ligera modificación relativa al dominio de aplicación. El conjunto de datos de entrada lo constituye de nuevo las notas de todas las ventanas de tiempo (por ejemplo 10 s) de todas las configuraciones de prueba. Pero en este caso, las notas se centran previamente en torno a una media general a fin de minimizar el efecto de discrepancias que ya se ha tratado en la primera etapa del proceso. A continuación se aplica el proceso habitual.

En primer lugar debe examinarse si esta distribución de notas para cada ventana de tiempo de cada configuración de prueba es «normal» o no, utilizando la prueba β_2. Si β_2 se encuentra entre 2 y 4 la distribución puede considerarse «normal». A continuación se aplica el proceso para cada ventana de tiempo de cada configuración de prueba como se expresa matemáticamente a continuación.

El primer paso del proceso es el cálculo de las notas centradas para cada ventana de tiempo y cada observador. La nota media, $\bar{u}_{klr}$, para cada configuración de prueba se define de la forma siguiente:

$$\bar{u}_{klr} = \frac{1}{N} \cdot \frac{1}{J} \sum_{n=1}^{N} \sum_{j=1}^{J} u_{njklr}$$

De forma similar, la nota media para cada configuración de prueba y cada observador se define así:

$$\bar{u}_{nklr} = \frac{1}{J} \sum_{j=1}^{J} u_{njklr}$$

y u_{njklr} corresponde a la nota del observador i para la ventana de tiempo j, la condición de tiempo k, la secuencia l y la repetición r.

Para cada observador, las notas centradas u^*_{njklr} se calculan de la forma siguiente:

$$u^*_{njklr} = u_{njklr} - \bar{u}_{nklr} + \bar{u}_{klr}$$

Para cada ventana de tiempo de cada configuración de prueba, se calcula la media, $\bar{u}^*_{jklr}$, la desviación típica, S^*_{jklr} y el coeficiente $\beta_2{}^*_{jklr}$, que viene dado por:

$$\beta_2{}^*_{jklr} = \frac{m_4}{(m_2)^2} \quad \text{con} \quad m_x = \frac{\sum_{n=1}^{N} (u^*_{njklr})^x}{N}$$

Para cada observador i, se determinan P^*_i y Q^*_i, es decir:

Para $j, k, l, r = 1, 1, 1, 1$, a J, K, L, R

Si $2 \le \beta_2{}^*_{jklr} \le 4$, entonces:

si $u^*_{njklr} \ge \bar{u}^*_{jklr} + 2\, S^*_{jklr}$ entonces $P^*_i = P^*_i + 1$

si $u^*_{njklr} \le \bar{u}^*_{jklr} - 2\, S^*_{jklr}$ entonces $Q^*_i = Q^*_i + 1$

o bien:

si $u^*_{njklr} \ge \bar{u}^*_{jklr} + \sqrt{20}\, S^*_{jklr}$ entonces $P^*_i = P^*_i + 1$

si $u^*_{njklr} \le \bar{u}^*_{jklr} - \sqrt{20}\, S^*_{jklr}$ entonces $Q^*_i = Q^*_i + 1$

Si $\dfrac{P^*_i + Q^*_i}{J \cdot K \cdot L \cdot R} > Y$ y $\left| \dfrac{P^*_i - Q^*_i}{P^*_i + Q^*_i} \right| < Z$ se rechaza al observador i

siendo:

N: número de observadores

J: número de ventanas de tiempo en una combinación de prueba de condición y secuencias de prueba

K: número de condiciones de prueba

L: número de secuencias

R: número de repeticiones.

Los valores propuestos para los parámetros (X, Y, Z) experimentados y adaptados a este método son: 0,2, 0,1, 0,3.

3 Procesamiento para encontrar una relación entre la nota media y la medición objetiva de la distorsión de imagen

Si las pruebas subjetivas se han realizado para determinar la relación entre la medición objetiva de una distorsión y las notas medias $\bar{u}$ ($\bar{u}$ calculado de acuerdo con el § 2.1), puede ser útil el siguiente proceso que consiste en encontrar una relación continua sencilla entre $\bar{u}$ y el parámetro de degradación.

3.1 Aproximación por una función logística simétrica

La aproximación de esta relación experimental por una función logística ofrece particular interés.

Las operaciones a que se someten los datos relativos a $\bar{u}$ pueden efectuarse de la manera siguiente:

La escala de valores de $\bar{u}$ se normaliza tomando una variable continua p, tal que:

$$p = (\bar{u} - u_{mín}) / (u_{máx} - u_{mín}) \tag{5}$$

siendo:

$u_{mín}$: nota mínima disponible en la escala u para la peor calidad

$u_{máx}$: nota máxima disponible en la escala u para la mejor calidad.

La representación gráfica de la relación entre p y D muestra que la curva tiende a presentar una forma sigmoide antisimétrica, siempre que los límites naturales de los valores de D, fuera de la región en que u varía rápidamente, sean lo suficientemente amplios.

La función $p = f(D)$ puede aproximarse entonces utilizando una función logística convenientemente elegida, tal como la que viene dada por la relación general siguiente:

$$p = 1/[1 + \exp (D - D_M) G] \tag{6}$$

donde D_M y G son constantes y G puede ser positivo o negativo.

El valor p, obtenido mediante la aproximación de la función logística óptima, se utiliza para hallar un valor numérico I tal que:

$$I = (1/p - 1) \tag{7}$$

Los valores de D_M y G pueden obtenerse a partir de datos experimentales mediante la siguiente transformación:

$$I = \exp (D - D_M) G \tag{8}$$

Utilizando una escala logarítmica para I se obtiene la relación lineal:

$$\log_e I = (D - D_M) G \tag{9}$$

La interpolación de una línea recta es sencilla y en algunos casos su precisión permite considerar que dicha línea recta representa la degradación debida al efecto medido por D.

La pendiente de la característica se expresa entonces mediante:

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \tag{10}$$

que proporciona el valor óptimo de G. D_M es el valor de D para $I = 1$.

La línea recta puede designarse por la característica de degradación asociada a la degradación específica que se considera. Se observará que la línea recta puede definirse por los valores característicos D_M y G de la función logística.

3.2 Aproximación por una función no simétrica

3.2.1 Descripción de la función

La aproximación de la relación entre las notas experimentales y la medición objetiva de una distorsión de imagen por una función logística simétrica tiene más éxito cuando el parámetro de distorsión *D* puede medirse en una unidad relacionada, por ejemplo la relación *S*/*N* (dB). Si el parámetro de distorsión se midió en una unidad física *d*, por ejemplo un retardo de tiempo (ms), la relación (8) debe sustituirse por la siguiente:

$$I = (d / d_M)^{1/G} \tag{11}$$

y, por consiguiente, la relación (6) pasa a ser:

$$p = 1 \Big/ \left[1 + (d / d_M)^{1/\mathrm{G}} \right] \tag{12}$$

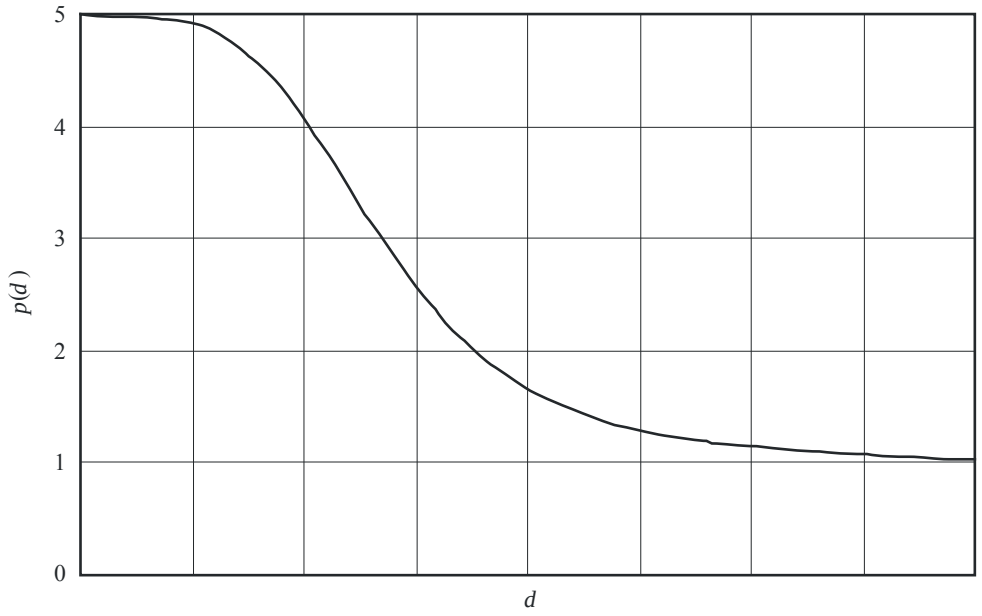
Esta función aproxima la función logística de una forma no simétrica.

3.2.2 Estimación de los parámetros de la aproximación

La estimación de los parámetros óptimos de la función que proporciona los errores mínimos residuales entre los datos reales y la función se puede obtener con cualquier algoritmo de estimación recurrente. La Fig. 16 muestra un ejemplo del uso de la función no simétrica para representar datos subjetivos reales. Esta representación permite estimar mediciones objetivas específicas correspondientes a un valor subjetivo interesante: 4,5 en la escala de cinco notas, por ejemplo.

FIGURA 16

Aproximación no simétrica



BT.0500-16

3.3 Corrección de la degradación/mejora residual y de los efectos de límite de escala

En la práctica, la utilización de una función logística a veces no puede evitar algunas diferencias entre los datos experimentales y la aproximación. Estas discrepancias pueden ser debidas a los efectos de fin de escala o a la presencia simultánea de varias degradaciones en la prueba que pueden repercutir en el modelo estadístico y deformar la función logística teórica.

Se ha identificado un tipo de efecto de límite de escala en el cual los observadores tienden a no utilizar los valores extremos de la escala de juicios, en particular para las notas de alta calidad. Ello puede deberse a un cierto número de factores, incluida la resistencia de tipo psicológico a realizar juicios extremos. Además, la utilización de la media aritmética de los juicios de acuerdo con la ecuación (1) cerca de los límites de la escala puede provocar resultados sesgados debido a la distribución no gaussiana de los votos en estas zonas.

Frecuentemente se indica en las pruebas una degradación residual (incluso en las imágenes de referencia la nota media alcanza únicamente un valor $\overline{u}_0 < u_{máx}$) .

Existen algunos mecanismos útiles para corregir los datos en bruto obtenidos de las evaluaciones a fin de lograr conclusiones válidas (véase el Cuadro 5).

La corrección de los efectos de límite, en caso de que existan en los datos experimentales, constituye una parte muy importante del procesamiento de datos. Por consiguiente, la elección del procedimiento debe efectuarse con un gran cuidado. Obsérvese que estos procedimientos de corrección suponen hipótesis especiales y, por consiguiente, es preciso tener precaución al utilizarlos; en la presentación de los resultados debe informarse que se han empleado dichos procedimientos.

CUADRO 5

Comparación de métodos de corrección de los efectos de límite de escala

Métodos de compensaciónde los efectos de límites	**Características**		
	Compensación de la degradación residual	Compensación de la mejora residual	Deriva en el centro de la escala
Sin compensación	No	No	No
Transformación de escala lineal	Sí	Puede ser un error significativo	No
Transformación de escala no lineal(1)	Sí	Sí	No
Método basado en la adición de degradaciones	Sí	No	Sí
Método multiplicativo	Sí	No	Sí

(1) De acuerdo con la transformación de escala no lineal deben calcularse los datos corregidos:

$$u_{corr} = C(\overline{u} - u_{mid}) + u_{mid}$$

$$C = \frac{\overline{u} - u_{0\,mín}}{u_{0\,máx} - u_{0\,mín}} \frac{u_{máx} - u_{mid}}{u_{0\,máx} - u_{mid}} + \frac{u_{0\,máx} - \overline{u}}{u_{0\,máx} - u_{0\,mín}} \frac{u_{mín} - u_{mid}}{u_{0\,mín} - u_{mid}}$$

siendo:

u_{corr}: nota corregida

$\overline{u}$: nota experimental sin corregir

$u_{mín}$, $u_{máx}$: límites de la escala de votación

u_{mid}: mitad de la escala de votación

$u0_{mín}$, $u0_{máx}$: límites inferior y superior de la tendencia de las notas experimentales.

3.4 Incorporación de los aspectos de fiabilidad a los gráficos

A partir de las notas medias de cada degradación sometida a prueba y del intervalo de confianza del 95% asociado, se elaboran tres series de notas:

– serie de notas mínimas (medias – intervalos de confianza);

– serie de notas medias;

– serie de notas máximas (medias + intervalos de confianza).

Se procede a continuación a una estimación de los parámetros independientemente para las tres series. Esto permite representar las tres funciones obtenidas en el mismo gráfico: las dos funciones derivadas de las series de notas máximas y mínimas en líneas de trazos, la estimación media en línea continua. Se señalan también en el gráfico los valores experimentales (véase la Fig. 17). Se obtiene así una estimación de la zona de confianza continua del 95%.

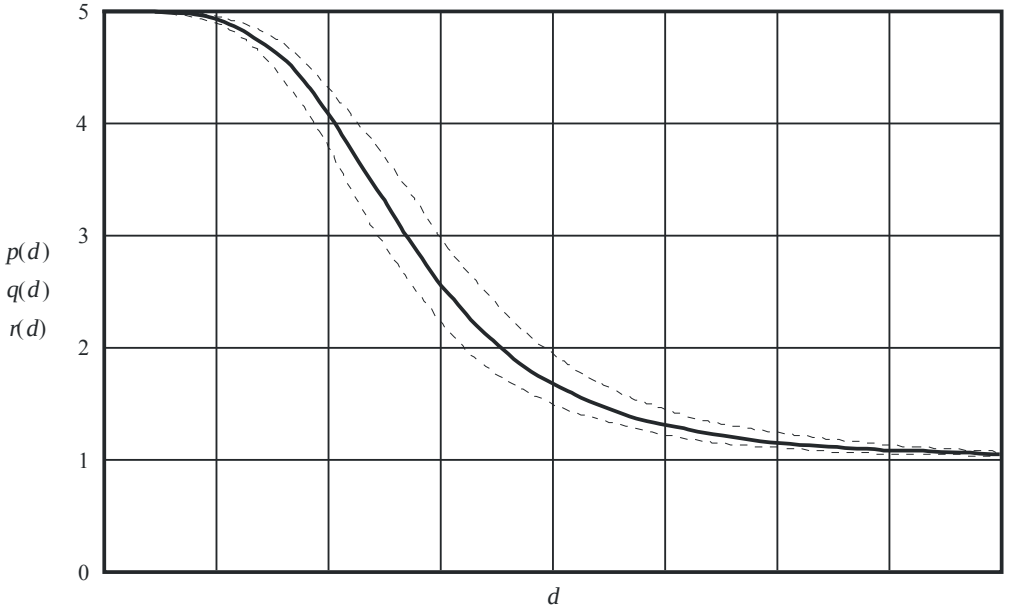
Con respecto a la nota 4,5 (umbral de visibilidad asociado al método), se obtiene directamente por lectura del gráfico un intervalo de confianza estimado del 95% que puede servir para determinar una gama de tolerancia.

La separación entre las curvas de máximas y mínimas no es un intervalo del 95%, sino una estimación media de éste.

Al menos el 95% de los valores experimentales debería estar incluido dentro de la zona de confianza; en caso contrario, podría pensarse que se ha producido un problema en la realización de la prueba o que el modelo de función elegido no es el óptimo.

FIGURA 17

Caso de una característica de degradación no simétrica



$p(d)$: series de notas medias
$q(d)$: series de notas mínimas
$r(d)$: series de notas máxima
d: medida objetiva de la degradación

BT.0500-17

4 Conclusiones

Se ha descrito un procedimiento para la evaluación de los intervalos de confianza, es decir, la precisión de un conjunto de pruebas de evaluación subjetiva.

El procedimiento permite también la estimación de magnitudes generales medias, que son aplicables no solamente al experimento particular que se está realizando, sino también a otros llevados a cabo según la misma metodología.

Por tanto, se pueden utilizar dichas magnitudes para dibujar diagramas del comportamiento del intervalo de confianza, que constituyen una ayuda tanto para las evaluaciones subjetivas como para la planificación de pruebas futuras.

Anexo 3

Descripción de un formato común para el intercambio de fichero

La finalidad de un formato común para el intercambio de fichero es facilitar el intercambio de datos entre laboratorios que participen en una campaña de evaluación subjetiva internacional en colaboración.

Una evaluación subjetiva se desarrolla en cinco fases sucesivas y dependientes entre sí: preparación de la prueba, realización de la prueba, procesamiento de los datos, presentación de los resultados e interpretación de los mismos. En grandes campañas internacionales, el trabajo se suele distribuir entre los diferentes laboratorios participantes:

– Un laboratorio se ocupa de la configuración de la prueba, en colaboración con otros participantes, identificando los parámetros de calidad que se han de evaluar, el material de la prueba que se ha de utilizar (en la actualidad, crítico pero no indebidamente crítico), el marco de la prueba (por ejemplo, metodología, distancia de observación, disposición de la sesión, secuencia de presentación de elementos de prueba) y el entorno de la prueba (por ejemplo, condiciones de observación, alocución introductoria).
– Se pide a los laboratorios que colaboran voluntariamente que proporcionen el material de prueba procesado de acuerdo con las técnicas adecuadas representativas del parámetro de calidad que se ha de evaluar (por simulación o en base a equipos físicos).
– Otro participante se encarga del montaje de la cinta de prueba.
– Diversos laboratorios colaboradores efectúan la prueba utilizando la cinta montada preliminar. La prueba puede ser una prueba ciega. En este caso, el laboratorio la llevará a cabo recogiendo los votos de los evaluadores sin tener que conocer necesariamente los parámetros de calidad objeto de evaluación.
– A otro participante se le pide generalmente que coordine la recogida de los datos brutos resultantes para procesamiento y publicación de los resultados, lo que también puede hacerse de manera ciega.
– Por último, se interpretan los resultados de un texto/cuadro o representación gráfica y se publica el informe final.

El formato propuesto permite reunir los resultados entregados de acuerdo con los procedimientos de prueba definidos durante la fase de definición de la prueba.

Este formato es conforme a los métodos de evaluación descritos en la Recomendación UIT-R BT.500.

Está constituido por ficheros de texto con la estructura que se muestra en los Cuadros 6 y 7. Su sintaxis se basa en etiquetas y campos y en un conjunto limitado de símbolos reservados (por ejemplo, «**[**», «**]**», « », «↵» y «=»).

No existe ninguna limitación intrínseca por lo que se refiere a capacidad (por ejemplo, el número de laboratorios participantes, observadores, secuencias de prueba y parámetros de calidad, límites de la escala de votación o tipo de periférico de votación).

CUADRO 6

Formato del fichero de texto Resultados de identificación

Formato y sintaxis del fichero de identificación	Comentarios
[Marco de la prueba]↵	[Identificador de sección]
Tipo = *«DSCQS» o «DSIS I», «DSIS II», etc.*↵	Identificación del método de la Recomendación UIT-R BT.500 utilizado
Número de sesiones = *1 ≤ entero ≤ x*↵	Número de sesiones(1) en las que se ha distribuido una prueba
Mínimo de la escala = *entero*↵	Definición de la escala (véanse los requisitos específicos del método, si existen)
Máximo de la escala = *entero*↵	
Tamaño del monitor = *entero*↵	Diagonal de la pantalla (pulgadas)
Marca y modelo del monitor = *cadena de caracteres*↵	
[RESULTADOS] ↵	[Identificador de sección]
Número de resultados = *1 ≤ entero ≤ y*↵	Número de ficheros Resultados(1) que se consideran
Resultado(*j*).Nombre de fichero(s) = *cadena de caracteres*.DAT↵	Nombre del fichero Completo.DAT (véase el Cuadro 7) incluyendo el trayecto
...	
Resultado(*j*).Nombre = *cadena de caracteres*↵	Nombre del fichero Resultados del cliente
Resultado(*j*).Laboratorio = *cadena de caracteres*↵	Identificación del laboratorio que efectúa la prueba
Resultado(*j*).Número de observadores = *1 ≤ entero ≤ N*↵	Número total de observadores
Resultado(*j*).*Entrenamiento* = *«Sí» o «No»*↵	Indica si los votos recogidos durante el entrenamiento se incluyen en el fichero DAT adjunto
[Resultado(*j*).Sesión (*i*).Observadores] ↵	[Identificador de sección]
O(*k*).Nombre = *cadena de caracteres*↵	Identificación del observador
O(*k*).Apellido = *cadena de caracteres*↵	
O(*k*).Sexo = *«M» o «F»*↵	Opcional
O(*k*).Edad = *entero*↵	Opcional
O(*k*).Ocupación = *cadena de caracteres*↵	Principales grupos socioeconómicos (por ejemplo, trabajador, estudiante)
O(*k*).Distancia = *entero*↵	Distancia de observación en alturas de la pantalla (por ejemplo, 3 *H*, 4 *H*, 6 *H*)

(1) Sesión: Una prueba se puede dividir en varias secciones diferentes para cumplir el requisito de duración de prueba máxima. El mismo observador u observadores diferentes pueden participar en distintas sesiones durante las cuales se les pedirá que evalúen configuraciones diferentes. Reuniendo los votos recogidos durante las distintas sesiones se obtiene un conjunto completo de Resultados (número de presentaciones × número de votos por presentación) de la prueba. Se puede adjuntar Resultados a los diversos ficheros .DAT que se entregarán por cada realización de prueba.

CUADRO 7

Formato del fichero de texto de datos brutos Resultados.DAT

Formato y sintaxis del fichero nombre de fichero .DAT	Comentarios
entero entero entero..... ↲ entero entero entero..... ↲	Un fichero de datos brutos DAT se compone de valores de votos separados por un espacio. Se ha de utilizar una línea por observador
entero entero entero..... ↲	Los datos brutos se almacenan según su orden de entrada
.....	Los datos se pueden distribuir en diferentes ficheros DAT identificados en el Cuadro 6 por Resultado(*j*). Nombre de fichero(s)[(1)]

(1) Véase la llamada (1) del Cuadro 6.