ITUPublicaciones
Recomendaciones

Unión Internacional de Telecomunicaciones
Sector de Radiocomunicaciones

Recomendación UIT-R BT.500-15

(05/2023)

Serie BT: Servicio de radiodifusión (televisión)

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

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 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 https://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, 2023

© UIT 2023

Reservados todos los derechos. Ninguna parte de esta publicación puede reproducirse por ningún procedimiento sin previa autorización escrita por parte de la UIT.

RECOMENDACIÓN UIT-R BT.500-15*

Metodologías para la evaluación subjetiva de la calidad[1] de las imágenes de televisión

(Cuestión UIT-R 102-4/6)

(1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002-2009-2012-2019-2023)

Cometido

En la presente Recomendación se describen metodologías para evaluar la calidad de las imágenes, incluidos los métodos generales de prueba, las escalas de apreciación utilizadas durante las evaluaciones y las condiciones de observación recomendadas para la realización de dichas evaluaciones. La Recomendación consta de tres partes:

– En la Parte 1 se describen los requisitos generales para la evaluación subjetiva de la calidad de las imágenes de televisión y se ofrece orientación sobre las circunstancias en que pueden utilizarse unas y otras metodologías.

– En la Parte 2 se describen las diversas metodologías de evaluación recomendadas que pueden utilizarse para la evaluación subjetiva de la calidad de las imágenes.

– En la Parte 3 se describen metodologías específicas para una serie de formatos de imagen y aplicaciones, basadas en las especificaciones que figuran en las Partes 1 y 2.

Palabras clave

Evaluación de imágenes, evaluación subjetiva

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 diferentes laboratorios en relación con varios de los aspectos que integran las metodologías de prueba subjetivas;

c) que la adopción de metodologías de evaluación normalizadas reviste una importancia particular 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 las metodologías recomendadas para la evaluación en laboratorio;

e) que la constante introducción de nuevas señales de televisión, métodos de procesamiento de señales y servicios de televisión novedosos o mejorados podría requerir la aplicación de distintas metodologías para la evaluación subjetiva de la calidad de las imágenes;

---

* La Comisión de Estudio 6 de Radiocomunicaciones introdujo modificaciones formales en la presente Recomendación en septiembre de 2023 y mayo de 2024, conforme a los estipulado en la Resolución UIT-R 1.

[1] Esta Recomendación debe ponerse en conocimiento de la Comisión de Estudio 12 del UIT-T.

f) que la introducción de dichos métodos de procesamiento, señales y servicios aumentará la probabilidad de que la calidad de funcionamiento de las distintas secciones de la cadena de la señal dependa cada vez más de procesos realizados en partes anteriores de la misma,

recomienda

1 que las metodologías generales de prueba, las escalas de apreciación y las condiciones de observación para la evaluación de la calidad de las imágenes descritas en la Parte 1 se utilicen para los experimentos de laboratorio y, siempre que sea posible, para las evaluaciones prácticas;

2 que, aunque existan metodologías alternativas y se elaboren otras nuevas, las descritas en la Parte 2 se utilicen siempre que proceda;

3 que las metodologías generales de prueba, las escalas de apreciación y las condiciones de observación para la evaluación de la calidad de las imágenes de un determinado sistema de imágenes o aplicación, que se describen en la Parte 3, se utilicen para los experimentos de laboratorio y, siempre que sea posible, para las evaluaciones prácticas;

4 que, para facilitar el intercambio de información entre los distintos laboratorios, se apliquen los requisitos de la metodología de prueba seleccionada, tal como se describe en la Parte 2;

5 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 Parte 1.

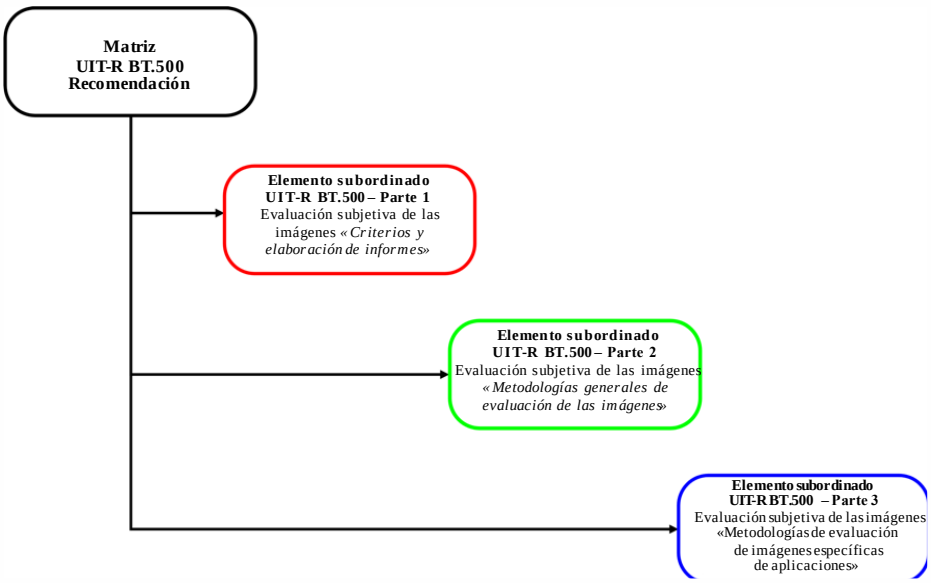
6 que, dada la importancia de sentar las bases de las evaluaciones subjetivas de las imágenes, todos los informes de pruebas contengan descripciones íntegras de las configuraciones y los materiales de prueba, los observadores y los métodos.

Notas relativas a la estructura y el uso de la presente Recomendación (a título informativo)

La Recomendación UIT-R BT.500 consta de tres Partes semiautónomas, subordinadas al presente texto matriz, según se indica en la Fig. 1.

FIGURA 1

Estructura de la Recomendación UIT-R BT.500



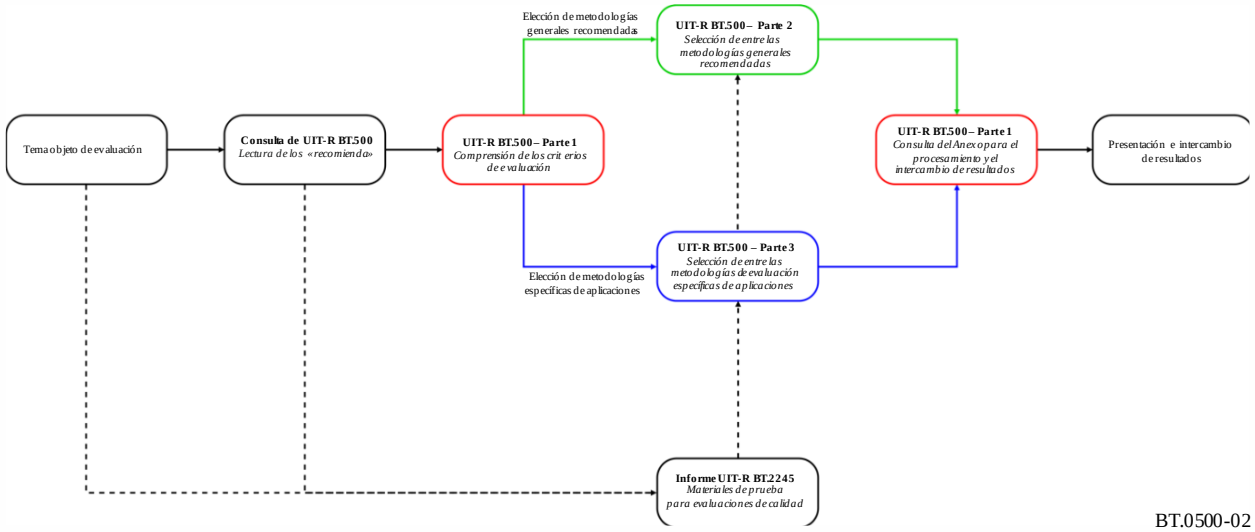
Se aconseja a los laboratorios que deseen realizar evaluaciones subjetivas de imágenes que consulten los *recomienda* anteriores y utilicen los criterios detallados en la Parte 1, a fin de discernir la metodología idónea para sus procedimientos de evaluación. En la Parte 2 se describen a grandes líneas varias metodologías recomendadas para la evaluación subjetiva de la calidad de las imágenes. En la Parte 3, se ofrece información sobre otras metodologías específicas de ciertas aplicaciones, que pueden facilitar la preparación de los procedimientos conexos de evaluación subjetiva.

Orientación relativa al uso de la Recomendación UIT-R BT.500

La Fig. 2 ilustra un posible cauce de utilización de la Recomendación UIT-R BT.500.

FIGURA 2

Utilización de la Recomendación UIT-R BT.500



Fundamentos

La estructura articulada en Partes de la presente versión de la Recomendación UIT-R BT.500 permite tanto añadir nuevas metodologías de evaluación subjetiva de las imágenes, como revisar las existentes, sin necesidad de elaborar nuevas Recomendaciones que repitan la información de múltiples documentos o publicar revisiones de partes que no requieren cambios.

Otras Recomendaciones relacionadas con la evaluación de imágenes

Las siguientes Recomendaciones versan sobre la medición objetiva de la calidad de las imágenes y pueden comprender otras metodologías de evaluación de imágenes específicas de ciertas aplicaciones, que utilizan parte de los criterios de evaluación de la Recomendación UIT-BT.500.

Recomendación UIT-R BT.1683	Técnicas de medición objetiva de la calidad de vídeo perceptual para la radiodifusión de televisión digital de definición convencional en presencia de una referencia completa
Recomendación UIT-R BT.1866	Técnicas de medición objetiva de la calidad visual percibida para aplicaciones de radiodifusión que utilizan la televisión de definición reducida en presencia de una señal de referencia completa
Recomendación UIT-R BT.1867	Técnicas de medición objetiva de la calidad visual percibida para aplicaciones de radiodifusión que utilizan la televisión de definición reducida en presencia de una referencia de anchura de banda reducida

Recomendación UIT-R BT.1885	Técnicas de medición objetiva de la calidad de vídeo perceptual para la radiodifusión de televisión digital de definición convencional en presencia de una referencia de anchura de banda reducida
Recomendación UIT-R BT.1907	Técnicas de medición objetiva de la calidad de percepción de vídeo para las aplicaciones de radiodifusión que utilizan TVAD en presencia de una señal de referencia completa
Recomendación UIT-R BT.1908	Técnicas de evaluación de la calidad de vídeo objetiva para las aplicaciones de radiodifusión que utilizan TVAD en presencia de una señal de referencia reducida

PARTE 1

Resumen de los requisitos de las evaluaciones subjetivas de las imágenes

ÍNDICE

Página

PARTE 1 5

Resumen de los requisitos de las evaluaciones subjetivas de las imágenes 5

1 Introducción 7

2 Características de evaluación comunes 7

2.1 Condiciones generales de observación 7

2.1.1 Condiciones generales de observación para efectuar evaluaciones subjetivas en un entorno de laboratorio 8

2.1.2 Condiciones generales de observación para efectuar evaluaciones subjetivas en el entorno doméstico 8

2.1.3 Distancia de observación 8

2.1.4 Ángulo de observación 10

2.1.5 Gama cromática ambiental de la sala 10

2.1.6 La pantalla 10

2.2 Señales fuente 12

2.3 Selección del material de prueba 13

2.3.1 Secuencias de prueba del UIT-R 13

2.4 Gama de condiciones y anclaje 14

2.5 Observadores 14

2.5.1 Número de observadores 14

2.5.2 Proceso de selección de observadores 14

2.5.3 Instrucciones para la evaluación 15

2.6 Sesión de prueba 15

2.7 Presentación de los resultados 15

3 Selección del método de prueba 16

Anexo 1 a la Parte 1 – Análisis y presentación de los resultados 16

A1-1 Introducción 16

A1-2 Métodos comunes de análisis 16

Página

A1-2.1 Cálculo de notas medias 17

A1-2.2 Cálculo del intervalo de confianza 17

A1-2.3 Selección posterior de los observadores 18

A1-2.4 Cálculo de notas medias e intervalos de confianza en condiciones de prueba difíciles 23

A1-3 Procesamiento para encontrar una relación entre la nota media y la medición objetiva de una distorsión de imagen 24

A1-3.1 Aproximación por una función logística simétrica 25

A1-3.2 Aproximación por una función no simétrica 25

A1-3.3 Corrección de la degradación/mejora residual y de los efectos de límite de escala 26

A1-3.4 Incorporación de los aspectos de fiabilidad a los gráficos 27

A1-4 Conclusiones 28

Adjunto 1 al Anexo 1 – Implementación de referencia del método descrito en el § A1-2.4 ... 29

Anexo 2 a la Parte 1 – Descripción de un formato común para el intercambio de fichero 34

Anexo 3 (informativo) a la Parte 1 – Característica de fallo de la imagen según su contenido 36

A3-1 Introducción 36

A3-2 Obtención de la característica de fallo 36

A3-3 Utilización de la característica de fallo 38

Anexo 4 (informativo) a la Parte 1 – Método para determinar una característica de fallo compuesta para contenido de programa y condiciones de transmisión 39

A4-1 Introducción 39

A4-2 Análisis del contenido de programa 39

A4-3 Análisis del canal de transmisión 40

A4-4 Obtención de las características de fallo compuestas 40

Anexo 5 (informativo) a la Parte 1 – Efecto contextual 41

Anexo 6 (informativo) a la Parte 1 – Mediciones de información espacial y temporal 42

Anexo 7 (informativo) a la Parte 1 – Términos y definiciones 43

1 Introducción

Se utilizan métodos de evaluación subjetiva de las imágenes 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 objeto de prueba. 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 se hallan las evaluaciones encaminadas a determinar la calidad de funcionamiento de los sistemas en condiciones óptimas, las cuales suelen denominarse «evaluaciones de calidad». En segundo lugar se hallan las evaluaciones encaminadas a determinar la capacidad de los sistemas de mantener la calidad en condiciones no óptimas relacionadas con la transmisión o la emisión, las cuales suelen denominarse «evaluaciones de degradación».

Con objeto de efectuar las evaluaciones subjetivas idóneas, en primer lugar es necesario seleccionar, de entre las diferentes opciones disponibles, la metodología que mejor se adapte a las circunstancias y los objetivos específicos de la correspondiente evaluación.

Para facilitar la elección, deberían considerarse las características generales descritas en el § 2, con miras a discernir las opciones más adecuadas en relación con el problema o proceso objeto de evaluación.

Una vez comprendidas las opciones, en el § 3 de la Parte 1 se ofrece un resumen de las metodologías recomendadas para la evaluación de las imágenes, que puede facilitar la elección de la metodología idónea para el problema o proceso objeto de evaluación, habida cuenta del tipo de evaluador por el que se opte y de las circunstancias del entorno de evaluación.

No obstante, la elección de la metodología más adecuada depende de los objetivos de servicio que deba alcanzar el sistema sometido a prueba. En consecuencia, los procedimientos íntegros de evaluación de ciertas aplicaciones figuran en la Parte 2 y en otras Recomendaciones UIT-R.

2 Características de evaluación comunes

A continuación, 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 metodologías conexas.

NOTA – Al efectuar evaluaciones subjetivas de imágenes de elevada gama dinámica, es aconsejable consultar los documentos adicionales a los que se hace referencia, de estar disponibles, en las secciones correspondientes[2].

2.1 Condiciones generales 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 doméstico. Estos parámetros se han

---

[2] Esta Recomendación será objeto de revisión, a medida que se vayan adquiriendo más conocimientos teóricos y prácticos sobre las imágenes de elevada gama dinámica, con miras a la inclusión de orientaciones adicionales.

seleccionado para definir un entorno ligeramente más crítico que las situaciones normales de observación en los hogares.

2.1.1 Condiciones generales de observación para efectuar evaluaciones subjetivas en un entorno de laboratorio

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

a)	Iluminación de la sala:	baja
b)	Cromaticidad del fondo:	D_{65}
c)	Luminancia de cresta[3]:	70-250 cd/m^2 (véase el § 2.1.6.5)
d)	Relación de contraste de la pantalla:	≤ 0,02 (véase el § 2.1.6.4)
e)	Relación entre la luminancia de fondo detrás de la pantalla de imágenes y el valor de cresta de luminancia de la imagen:	≈ 0,15

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

a)	Luminancia del medio ambiente en la pantalla (la luz incidente del entorno proyectada sobre la pantalla debe medirse perpendicularmente a la misma):	200 lux
b)	Luminación de cresta:	70-500 cd/m^2 (véase el § 2.1.6.4)
c)	Relación entre la luminancia de pantalla inactiva y el valor de cresta de la luminancia de pantalla):	≤ 0,02 (véase el § 2.1.6.4)

2.1.3 Distancia de observación

La distancia de observación se basa en el tamaño de la pantalla y puede elegirse según dos criterios distintos: la distancia de observación preferida (PVD, *preferred viewing distance*) y la distancia de observación de diseño (DVD, *design viewing distance*). La selección de uno u otro criterio dependerá del objeto del estudio.

---

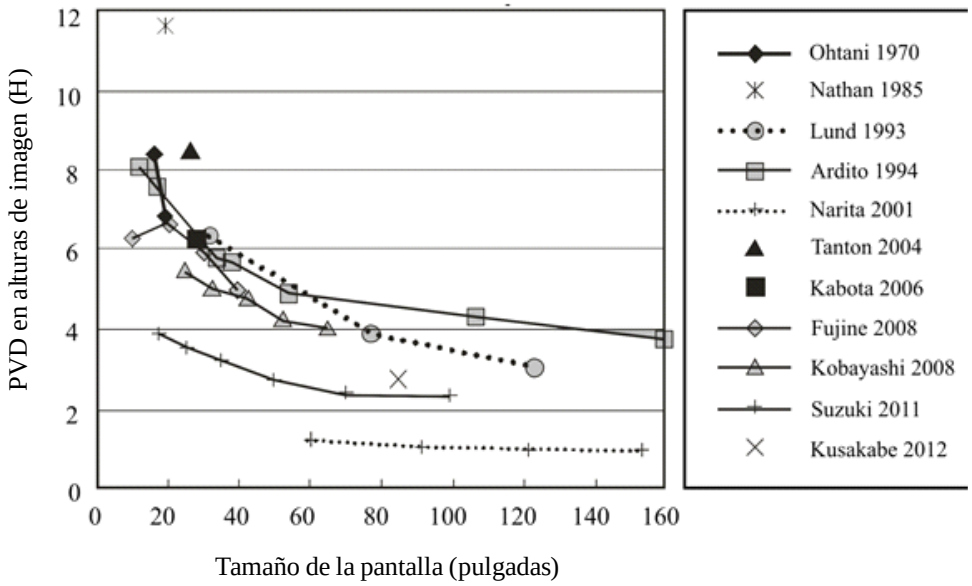
[3] La luminación de cresta debe ajustarse de acuerdo con la iluminación de la sala.

2.1.3.1 Distancia de observación preferida

La distancia de observación preferida (PVD) se basa en las preferencias de los usuarios que se han determinado empíricamente. La PVD (en función de los tamaños de pantalla) se presenta en la Fig. 1-1, que contiene un cierto número de conjuntos de datos recopilados de las fuentes disponibles. Esta información puede utilizarse para diseñar una prueba de evaluación subjetiva.

FIGURA 1-1

Distancia de observación preferida en función de los tamaños de pantalla



BT.0500-0 1-1

2.1.3.2 Distancia de observación de diseño

La distancia de observación de diseño (DVD), o distancia de observación óptima, en un sistema digital es la distancia a la cual dos píxeles adyacentes subtienden un ángulo de 1 arc-minuto en el ojo del observador; y el ángulo de observación horizontal óptimo es el ángulo bajo el cual se ve una imagen a su distancia de observación óptima.

El Cuadro 1-1 presenta las distancias de observación óptimas (y los ángulos de observación horizontal óptimos) para varios sistemas de resolución de imagen, expresadas en múltiplos de la altura de la imagen.

CUADRO 1-1

Ángulo de observación horizontal óptimo, distancia de observación óptima en alturas de imagen (H)

Sistema de imagen	Referencia	Formato de imagen	Formato de imagen del píxel	Ángulo de observación horizontal óptimo	Distancia de observación óptima
720 × 483	UIT-R BT.601	4:3	0,89	11°	7 *H*
640 × 480	VGA	4:3	1	11°	7 *H*
720 × 576	UIT-R BT.601	4:3	1,07	13°	6 *H*
1 024 × 768	XGA	4:3	1	17°	4,5 *H*
1 280 × 720	UIT-R BT.1543 y BT.1874	16:9	1	21°	4,8 *H*
1 400 × 1 050	SXGA+	4:3	1	23°	3,3 *H*
1 920 × 1 080	UIT-R BT.709	16:9	1	31°	3,2 *H*
3 840 × 2 160	UIT-R BT.2020	16:9	1	58°	1,6 *H*
7 680 × 4 320	UIT-R BT.2020	16:9	1	96°	0,8 *H*

NOTA – Cuando la evaluación de la imagen abarca la resolución, conviene utilizar el valor de distancia de observación más bajo para los formatos 7 680 × 4 320 y 3 840 × 2 160. En caso contrario, podrá optarse por cualquier distancia en la gama indicada (para el formato 3 840 × 2 160: de 1,6 a 3,2 veces la altura de imagen; para el formato 7 680 × 4 320: de 0,8 a 3,2 veces la altura de imagen).

2.1.4 Ángulo de observación

El máximo ángulo de observación con respecto al ángulo de observación normal debe limitarse de forma que las desviaciones en el color reproducido en la pantalla no sean visibles al observador. Para determinar el ángulo de observación, también debe considerarse el ángulo de observación horizontal óptimo de un sistema de imágenes sometido a prueba. A fin de obtener más información al respecto, véase el § 1.8 del Informe UIT-R BT.2129.

2.1.5 Gama cromática ambiental de la sala

El fondo de la pantalla debe ser del mismo color que el punto blanco de referencia; para el resto de la sala deben utilizarse superficies mates oscuras. El objetivo es minimizar la luz dispersa en la pantalla.

2.1.6 La pantalla

Al utilizar pantallas con características distintas, se obtienen diferentes valores subjetivos de calidad de la imagen. Por consiguiente, se recomienda encarecidamente verificar de antemano las características de las pantallas utilizadas. En los casos en que se utilicen pantallas planas profesionales para la evaluación subjetiva, pueden consultarse tanto la Recomendación UIT-R BT.1886, «Función de transferencia electroóptica de referencia para las pantallas planas utilizadas en la producción de TVAD en estudio», como el Informe UIT-R BT.2129, «Requisitos de usuario para un monitor de pantalla plana empleado como monitor principal en un entorno de producción de programas de TVAD».

El Informe UIT-R BT.2390 contiene información relativa a las pantallas domésticas y de laboratorio, así como a las condiciones de observación, para la evaluación de imágenes de elevada gama dinámica (HDR).

2.1.6.1 Procesamiento en la pantalla

En su caso, las técnicas de procesamiento en la pantalla, véanse cambios de escala de la imagen, conversiones de velocidad de trama o potenciaciones de imagen, deben aplicarse de forma que no se introduzcan perturbaciones. Cabe optar por el procesamiento HDR adecuado para el sistema HDR evaluado o utilizado durante la evaluación. En las evaluaciones relacionadas con el entorno del consumidor o la distribución, ello puede incluir el uso de los correspondientes metadatos estáticos o dinámicos. En las notas de las evaluaciones debe incluirse información detallada sobre esos metadatos, para que otros laboratorios puedan repetir con exactitud las evaluaciones.

Cuando se utilicen pantallas domésticas para la evaluación subjetiva de la calidad de las imágenes, es importante desactivar todas las opciones de procesamiento de imágenes, a menos que la repercusión de dicho procesamiento sea el objeto de la evaluación.

En los casos en que se acceda a imágenes entrelazadas, cabría indicar en el informe de prueba si se ha utilizado el desentrelazador o no. Si las señales entrelazadas pueden visualizarse sin el desentrelazador, es preferible no utilizarlo.

2.1.6.2 Resolución de la pantalla

Normalmente, la resolución de las pantallas profesionales satisface las normas requeridas para realizar evaluaciones subjetivas en su gama de funcionamiento de luminancia.

Para verificar y notificar 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 se utilizan aparatos de TV domésticos con pantalla plana para efectuar las evaluaciones subjetivas, se recomienda encarecidamente verificar y notificar 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 las resoluciones de las pantallas o de los aparatos de televisión domésticos, utilizan un diagrama de prueba de barrido generado electrónicamente.

2.1.6.3 Ajuste de la pantalla

El brillo y el contraste de la pantalla deben ajustarse con arreglo a la luminancia del entorno, utilizando las formas de onda PLUGE, de conformidad con la Recomendación UIT-R BT.814.

A efectos de la evaluación de imágenes de gama dinámica convencional (SDR), la relación de contraste de la pantalla debe medirse de acuerdo con la Recomendación UIT R BT.815. En los casos en que se evalúen imágenes HDR, debería consultarse el Informe UIT-R BT.2390.

2.1.6.4 Contraste de la pantalla

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

Las pantallas profesionales raramente utilizan tecnologías para mejorar su contraste en un entorno de alta luminancia, por lo que es posible que no cumplan la norma de contraste necesaria si se utilizan en dichos entornos.

Las pantallas domésticas suelen emplear distintas tecnologías para lograr un mejor contraste en un entorno de alta luminancia.

2.1.6.5 Brillo de la pantalla

Al ajustar el brillo de las pantallas LCD (pantallas de cristal líquido) es preferible utilizar un control de intensidad de la iluminación de fondo en vez de un nivel de señal para mantener la precisión de bit. En el caso de otras tecnologías de presentación que no utilizan iluminación de fondo, el nivel de blanco debe ajustarse por medio de mecanismos distintos al de escala de nivel de la señal. Obsérvese que el PDP (panel de visualización de plasma) controla el brillo mediante el número de radiaciones de luz y si se fija un valor de brillo menor se degradará la reproducción del tono.

2.1.6.6 Perturbaciones cinéticas

La pantalla no debe introducir perturbaciones cinéticas producidas por tecnologías de presentación específicas. Por otro lado, los efectos cinéticos, incluida la señal de entrada, deben representarse en la pantalla. En los casos en que se utilicen pantallas domésticas, es vital que TODAS las opciones de procesamiento cinético estén desactivadas.

2.1.6.7 Zonas de seguridad de los monitores con pantalla grande y formato de imagen 16:9

En la Recomendación UIT-R BT.1848 se indican las zonas de seguridad de los monitores con formato de imagen 16:9.

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 fijas y secuencias de vídeo almacenadas digitalmente son las fuentes de señales más reproducibles y, por consiguiente, las preferidas. Pueden intercambiarse entre laboratorios, para dar mayor significado a las comparaciones de sistemas.

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. En su caso, los analizadores de diapositivas de 35 mm pueden utilizarse como fuente de imágenes fijas. La resolución disponible 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.

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 1-2 se describen los problemas de evaluación y de materiales de prueba típicos utilizados para abordar esos problemas.

CUADRO 1-2

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 Anexos 3 y 4 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.

2.3.1 Secuencias de prueba del UIT-R

Varias organizaciones han desarrollado imágenes fijas y secuencias. El Informe UIT-R BT.2245, «Material de pruebas de la TVAD y la TVUAD para la evaluación de la calidad de la imagen», contiene información detallada sobre el material de pruebas de la TVAD y la TVUAD, que puede utilizarse para realizar evaluaciones subjetivas. En los Anexos 1 y 2 a la Parte 1 de la presente Recomendación 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.

2.5.1 Número de observadores

A menos que se indique otra cosa en la metodología elegida, deben emplearse al menos 15 observadores. El número de asesores necesarios depende de la sensibilidad y la fiabilidad del procedimiento de prueba adoptado y de la magnitud prevista 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.

2.5.2 Proceso de selección de observadores

Por lo general, 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).

En las secciones A1-2.3 y A1-2.4 se describen distintos procesos de selección de los observadores que pueden aplicarse a diversas metodologías de prueba. Cuando se realizan pruebas de laboratorio o de carácter menos formal en el marco de un programa de prueba en varias ubicaciones u organizaciones, es importante que se intercambien todos los detalles sobre el método y los criterios de selección de los observadores y que se adjunten a los resultados publicados.

Por regla general, se debe incluir el mayor número de detalles posible sobre las características de sus equipos de evaluación, entre los que podrían figurar la categoría profesional (por ejemplo, empleado de organización radiodifusora, estudiante de universidad, empleado de oficina...), sexo y edad.

NOTA – 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.

2.5.3 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.

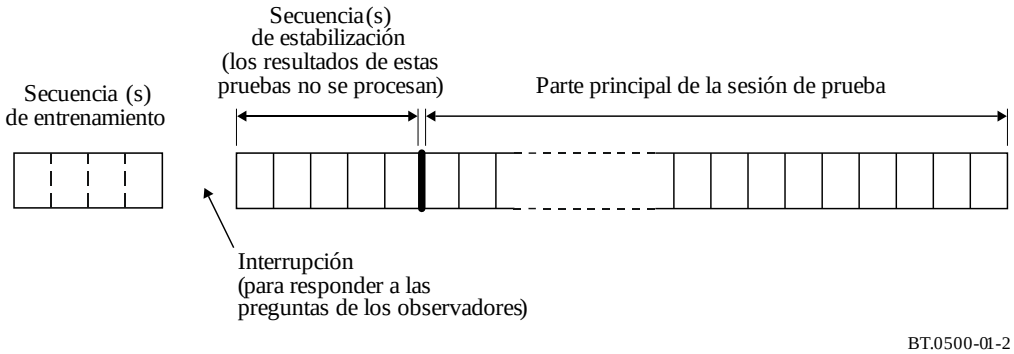
2.6 Sesión de prueba

Una sesión no debería durar más de 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 considerarse 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 1-2

Estructura de presentación de la sesión de prueba



2.7 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 1 a la Parte 1 de 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 imagen fuente y de pantallas (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 la Parte 3 de la presente Recomendación se ofrecen orientaciones para la evaluación subjetiva de la calidad de las imágenes en determinados formatos de imagen y aplicaciones.

Anexo 1
a la Parte 1

Análisis y presentación de los resultados

A1-1 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 que figuran en los Anexos 1, 2 y 3 a la Parte 2 de la presente Recomendación, así como 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.

A1-2 Métodos comunes de análisis

Las pruebas realizadas de acuerdo con los principios de los métodos descritos en la Parte 1 § 2 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.

A1-2.1 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.

A1-2.2 Cálculo del intervalo de confianza

A1-2.2.1 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:

$$\left[\bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr}\right] \tag{2}$$

donde:

$$\delta_{jkr} = 1{,}96 \frac{S_{jkr}}{\sqrt{N}} \tag{3}$$

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

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{4}$$

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.

A1-2.2.2 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.

A1-2.3 Selección posterior de los observadores

A1-2.3.1 Selección posterior basada en curtosis 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, $\bar{u}_{jkr}$, 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).

NOTA – 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, $\bar{u}_{klr}$, la desviación típica, S_{jkr}, y el coeficiente de curtosis, β_{2jkr}. Este coeficiente viene dado por:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{con} \quad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

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 $2 \leq \beta_{2jkr} \leq 4$, entonces:

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

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

o bien:

si $u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20}\;S_{jkr}$ entonces $P_i = P_i + 1$

si $u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20}\;S_{jkr}$ 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).

A1-2.3.2 Selección posterior basada en curtosis 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 $\bar{u}_{klr}$, la desviación típica, S_{jklr} y el coeficiente, β_{2jklr}. Este coeficiente viene dado por la expresión:

$$\beta_{2jklr} = \frac{m_4}{(m_2)^2} \quad \text{con} \quad m_x = \frac{\sum_{n=1}^{N} (u_{njklr} - \bar{u})^x}{N} \tag{6}$$

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 $2 \leq \beta_{2jklr} \leq 4$, entonces:

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

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

o bien:

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

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

Si $\dfrac{P_i}{J \cdot K \cdot L \cdot R} > X$ o $\dfrac{Q_i}{J \cdot K \cdot L \cdot R} > X$ 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. 1-3 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 presente Anexo. 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:

$$\bar{u}_{klr} = \frac{1}{N} \cdot \frac{1}{J} \sum_{n=1}^{N} \sum_{j=1}^{J} u_{njklr} \tag{7}$$

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

$$\bar{u}_{nklr} = \frac{1}{J} \sum_{j=1}^{J} u_{njklr} \tag{8}$$

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:

$$u^*_{njklr} = u_{njklr} - \bar{u}_{nklr} + \bar{u}_{klr} \tag{9}$$

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:

$$\beta_2{}^*_{jklr} = \frac{m_4}{(m_2)^2} \quad \text{con} \quad m_x = \frac{\sum_{n=1}^{N} (u^*_{njklr})^x}{N} \tag{10}$$

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.

A1-2.3.3 Selección posterior basada en correlación

Cada observador debe tener un método estable y coherente para votar una relativa degradación de calidad en cada escena y algoritmo. Los criterios de rechazo verifican el nivel de coherencia de las notas de un observador según la nota media de todos los observadores para una determinada sesión de pruebas. El criterio de decisión se basa en una correlación de notas individuales con respecto a las notas medias correspondientes de todos los observadores de la prueba. El procedimiento es más sencillo de implementar que el correspondiente descrito en las secciones anteriores.

A1-2.3.3.1 Correlación de Pearson

La relación entre la escala de calidad y la gama de notas de los observadores se supone lineal para aplicar la correlación de Pearson.

El principal objetivo es verificar mediante un método sencillo si las notas de un observador son coherentes con arreglo a las notas medias de todos los observadores en toda la prueba de la sesión. La referencia oculta se considera como un anclaje de alta calidad. Si se incluyen los anclajes bajos y altos, aumentan la nota de correlación, y a la inversa, los desplazamientos de correlación entre los observadores disminuyen.

$$r(x,y)=\frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \tag{11}$$

donde:

x_i: nota media de todos los observadores para el trío (algoritmo, velocidad binaria, escena)

y_i: nota individual de un observador para el mismo trio

n: (número de algoritmo) × (número de escenas)

i: {número de códec, número de velocidad binaria, número de escena}.

A1-2.3.3.2 Correlación de rango de Spearman

La correlación de rango de Spearman puede aplicarse incluso si la relación entre la escala de calidad y la gama de notas de los observadores no se supone lineal[4]:

$$r(x,y)=\left[1-\frac{6\times\sum_{i=1}^{n}\left[R(x_i)-R(y_i)\right]^2}{n^3-n}\right] \tag{12}$$

donde:

x_i: nota media de todos los observadores para el trío (algoritmo, velocidad binaria, escena)

y_i: nota individual de un observador para el mismo trío

n: (número de algoritmo) × (número de escenas)

$R(x_i$ o $y_i)$: es el orden del rango

i: {número de códec, número de velocidad binaria, número de escena}.

---

[4] Generalmente, los resultados de la correlación de Pearson están muy próximos a los de Spearman.

A1-2.3.3.3 Criterios finales de rechazo para descartar a un observador de una prueba

Las correlaciones de rango de Spearman y de Pearson se llevan a cabo para descartar uno o varios observadores en función de las condiciones siguientes:

SI [media(r) – sdt(r)] > Umbral de correlación máxima (MCT).
Umbral de rechazo = Umbral de correlación máxima (MCT).
EN OTRO CASO Umbral de rechazo = [media(r) – sdt(r)].

SI [r (observador$_i$)] > Umbral de rechazo.
ENTONCES observador «i» de la prueba no se descarta.
EN OTRO CASO observador «i» de la prueba se descarta.

donde:

r = min (correlación de Pearson, correlación de rango de Spearman)

media(r): promedio de las correlaciones de todos los observadores de una prueba

sdt(r): desviación típica de las correlaciones de todos los observadores de una prueba

Umbral de correlación máxima (MCT) = 0,85.

El valor 0,85 de MCT es válido para los métodos SAMVIQ y DSCQS, en caso contrario debe considerarse el valor 0,7 para los métodos SS y DSIS.

A1-2.4 Cálculo de notas medias e intervalos de confianza en condiciones de prueba difíciles

Muy a menudo, una prueba subjetiva debe realizarse en condiciones difíciles. Por ejemplo, en una prueba por colaboración masiva, los sujetos están expuestos a un entorno menos controlado que en un laboratorio. En una prueba a gran escala realizada por varios laboratorios, la variabilidad entre laboratorios podría dar lugar a una gran variación de las evaluaciones obtenidas. Los métodos presentados en los § A1-2.1 a A1-2.3 no suelen ser adecuados para tales circunstancias. En esta sección se presenta una técnica avanzada de análisis de datos que ha demostrado mejorar la calidad de los datos de las notas medias y los intervalos de confianza obtenidos. En el Adjunto 1 al presente Anexo también se puede encontrar una implementación de referencia en Python.

La idea que subyace a esta técnica es la siguiente. Conviene realizar un modelo explícito del comportamiento de cada sujeto; en particular, el sesgo y la coherencia de un sujeto son dos factores humanos destacados que afectan a los votos del sujeto. Mediante un procedimiento iterativo, esta técnica trata de estimar conjuntamente la calidad verdadera de cada presentación y el sesgo y la coherencia de cada sujeto. La calidad verdadera estimada de cada presentación puede interpretarse como una «nota media de opinión con ponderación de coherencia y sesgo eliminado». En comparación con la selección posterior de sujetos descrita en el § A1-2.3.1, en la que se mantienen o rechazan todos los votos de un sujeto («rechazo duro»), esta técnica puede describirse como «rechazo suave». Es decir, los votos de un sujeto atípico que vota de forma incoherente tendrían una ponderación pequeña, por lo que contribuirían poco a la MOS global. Una variante de esta técnica es la estimación del sesgo y la coherencia de cada sujeto. Se trata de información valiosa para determinar la idoneidad de un sujeto para realizar pruebas subjetivas, por lo que puede utilizarse para seleccionar sujetos para futuras pruebas. Por ejemplo, si un sujeto ha demostrado votar de forma muy incoherente, puede ser excluido de futuras sesiones.

Mediante esta técnica se estiman en primer lugar las notas medias correspondientes a cada una de las presentaciones considerando todos los sujetos y repeticiones:

$$\bar{u}_{jk} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} u_{ijkr} \tag{13}$$

donde u_{ijkr} es la nota del observador *i* para la condición *j*, secuencia/imagen *k*, repetición *r*, *N* es el número de observadores, y *R* representa el número de repeticiones.

La segunda etapa consiste en estimar el sesgo de cada observador b_i mediante la fórmula:

$$b_i = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} u_{ijkr} - \bar{u}_{jk} \tag{14}$$

siendo J y K el número de condiciones y el número de secuencias, respectivamente. Las siguientes etapas se llevan a cabo en un bucle iterativo.

La estimación actual de la nota media de cada presentación se registra como $\bar{u}_{jk}^{c}$, esto es,

$$\bar{u}_{jk}^{c} = \bar{u}_{jk} \tag{15}$$

y, a continuación, se calcula el residuo en cada calificación observada que no puede explicarse por la nota media y el sesgo del observador:

$$e_{ijkr} = u_{ijkr} - \bar{u}_{jk} - b_i \tag{16}$$

Estos residuos se utilizan después para calcular la incoherencia de cada observador σ_i de la forma siguiente:

$$\sigma_i = \sqrt{\frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} (u_{ijkr} - \mu_{e_i})^2} \tag{17}$$

donde:

$$\mu_{e_i} = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} e_{ijkr} \tag{18}$$

Las nuevas estimaciones de las notas medias pueden obtenerse entonces mediante:

$$\bar{u}_{jk} = \frac{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2} (u_{ijkr} - b_i)}{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2}} \tag{19}$$

y, a continuación, se actualiza el sesgo según la ecuación (12).

El bucle termina si:

$$\sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{u}_{jk} - \bar{u}_{jk}^{c})^2 \tag{20}$$

Tras la terminación, la desviación típica de la nota de cada presentación se obtiene de la forma siguiente:

$$S_{jk} = \frac{\sigma_j}{\sqrt{N}} \tag{21}$$

donde:

$$\sigma_j = \sqrt{\frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} (e_{ijkr} - \mu_{e_j})^2} \tag{22}$$

y

$$\mu_{e_j} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} e_{ijkr} \tag{23}$$

A continuación, se calcula el intervalo de confianza final según las ecuaciones (2) y (3).

A1-3 Procesamiento para encontrar una relación entre la nota media y la medición objetiva de una 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 § A1-2.1), puede ser útil el siguiente proceso que consiste en encontrar una relación continua sencilla entre $\bar{u}$ y el parámetro de degradación.

A1-3.1 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:

$$p = (\bar{u} - u_{mín})/(u_{máx} - u_{mín}) \tag{24}$$

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:

$$p = 1/[1 + \exp(D - D_M) \cdot G] \tag{25}$$

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:

$$I = (1/p - 1) \tag{26}$$

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

$$I = \exp(D - D_M) \cdot G \tag{27}$$

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

$$\log_e I = (D - D_M) \cdot G \tag{28}$$

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:

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \tag{29}$$

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.

A1-3.2 Aproximación por una función no simétrica

A1-3.2.1 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 (27) debe sustituirse por la siguiente:

$$I = (d/d_M)^{1/G} \tag{30}$$

y, por consiguiente, la ecuación (25) pasa a ser:

$$p = 1/\left[1 + (d/d_M)^{1/G}\right] \tag{31}$$

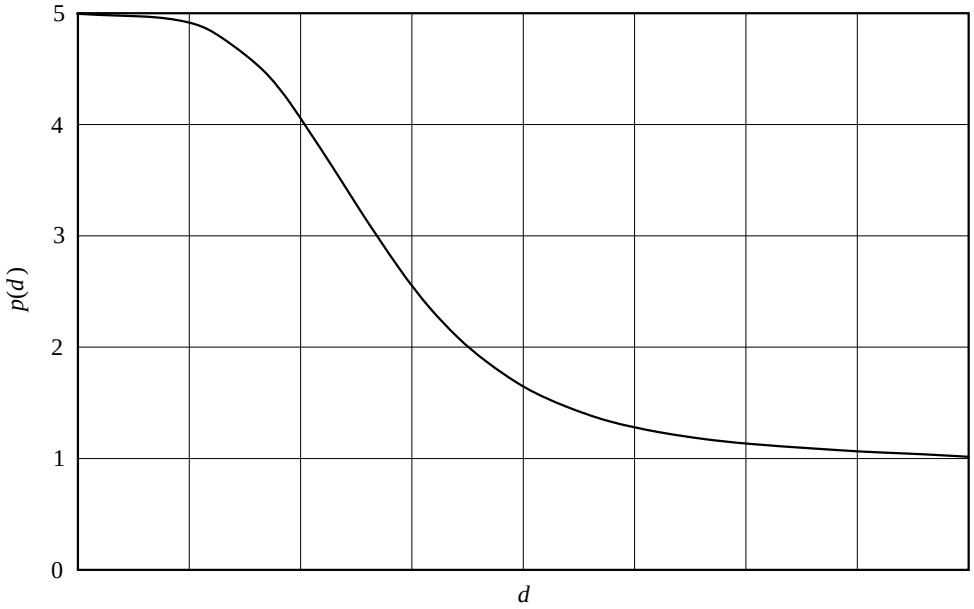
Esta función aproxima la función logística de una forma no simétrica.

A1-3.2.2 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. 1-3 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 1-3

Aproximación no simétrica



BT.0500-01-3

A1-3.3 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 $\bar{u}_0 < u_{máx}$).

Existen algunos mecanismos útiles para corregir los datos en bruto obtenidos de las evaluaciones a fin de lograr conclusiones válidas (véase el Cuadro 1-3).

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 1-3

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

Métodos de compensación de 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:

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

$$C = \frac{\bar{u} - u_{0_{min}}}{u_{0_{max}} - u_{0_{min}}} \frac{u_{max} - u_{mid}}{u_{0_{max}} - u_{mid}} + \frac{u_{0_{max}} - \bar{u}}{u_{0_{max}} - u_{0_{min}}} \frac{u_{min} - u_{mid}}{u_{0_{min}} - u_{mid}}$$

siendo:

u_{corr}: nota corregida

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

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

u_{mid}: mitad de la escala de votación

$u_{0mín}$, $u_{0máx}$: límites inferior y superior de la tendencia de las notas experimentales.

A1-3.4 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. 1-4). Se obtiene así una estimación de la zona de confianza continua del 95%.

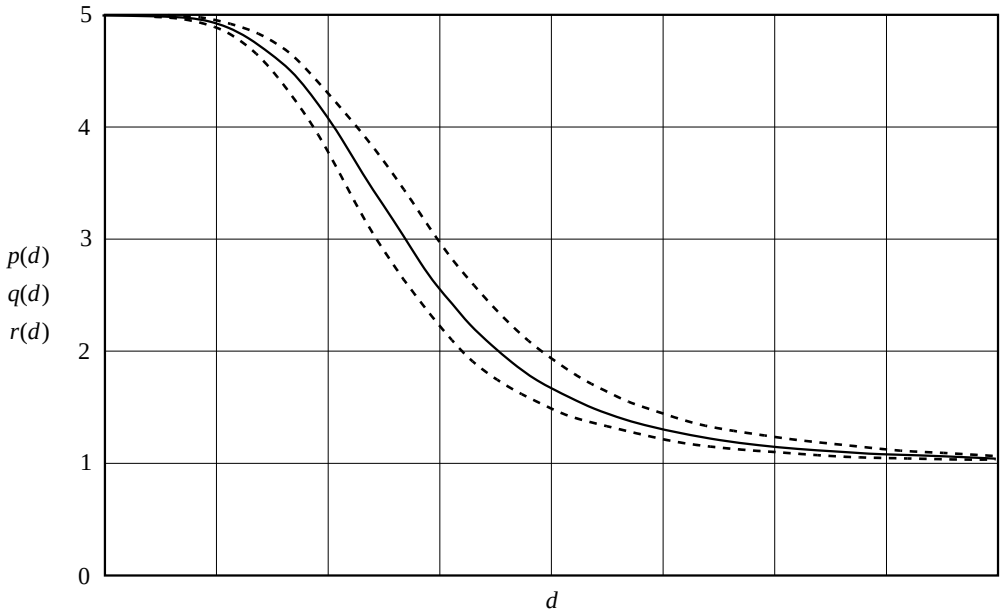
Con respecto a la nota 4,5 (umbral de visibilidad asociado al método), se obtiene directamente por lectura del gráfico un intervalo de confianza estimado del 95% que puede servir para determinar una gama de tolerancia.

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 1-4

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 máximas
d: medida objetiva de la degradación

BT.0500-01-4

A1-4 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.

Adjunto 1
al Anexo 1

Implementación de referencia del método descrito en el § A1-2.4

El presente Adjunto incluye una implementación de referencia en Python de la técnica de análisis de datos presentada en el § A1-2.4. El código y los datos de muestra utilizados también están a disposición del público en el paquete de Python SUREAL en la siguiente dirección: https://github.com/Netflix/sureal/tree/master/itur_bt500_demo.

Los datos de entrada se preparan de la siguiente manera. Los votos brutos se organizan en una matriz 2D, separados por comas. Cada fila corresponde a una presentación (imagen fuente en una condición de prueba); cada columna corresponde a un sujeto.

No es necesario que todos los sujetos voten en todas las presentaciones. Si el sujeto *i* no votó en una presentación *jk*, se introduce «nan» (no un número) en la posición (*jk*,*i*). Los datos de entrada se introducen en un archivo .csv. A continuación, se muestra un pequeño archivo .csv de muestra correspondiente a los votos de 20 sujetos y a 30 presentaciones con dos repeticiones.

```
5.0,nan,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,5.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
3.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,3.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,4.0,3.0,4.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
4.0,5.0,nan,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,4.0,5.0,1.0,4.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,4.0,3.0,5.0
3.0,5.0,4.0,2.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0
5.0,2.0,1.0,3.0,3.0,4.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,1.0,3.0,1.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,1.0,2.0
5.0,5.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,3.0,2.0,3.0,4.0,2.0,1.0,2.0,2.0,1.0,2.0,2.0
5.0,2.0,4.0,3.0,4.0,2.0,2.0,2.0,2.0,4.0,3.0,3.0,3.0,5.0,2.0,2.0,2.0,4.0,2.0,2.0
5.0,5.0,5.0,5.0,4.0,3.0,3.0,3.0,3.0,5.0,3.0,4.0,4.0,3.0,2.0,2.0,3.0,3.0,3.0,3.0
5.0,5.0,4.0,3.0,5.0,4.0,4.0,4.0,4.0,5.0,4.0,4.0,5.0,4.0,3.0,3.0,4.0,3.0,3.0,4.0
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
1.0,4.0,1.0,4.0,3.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,4.0
4.0,2.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
2.0,5.0,3.0,2.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
5.0,5.0,5.0,5.0,3.0,3.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0
4.0,5.0,5.0,3.0,5.0,2.0,2.0,3.0,1.0,3.0,3.0,2.0,3.0,5.0,1.0,1.0,2.0,2.0,2.0,2.0
1.0,2.0,2.0,4.0,5.0,1.0,2.0,2.0,1.0,3.0,2.0,2.0,4.0,2.0,3.0,1.0,2.0,2.0,1.0,3.0
4.0,5.0,3.0,5.0,2.0,3.0,2.0,3.0,3.0,4.0,2.0,3.0,4.0,3.0,3.0,1.0,2.0,2.0,2.0,3.0
1.0,5.0,3.0,5.0,4.0,2.0,3.0,3.0,3.0,5.0,3.0,3.0,4.0,2.0,3.0,2.0,3.0,3.0,2.0,3.0
5.0,5.0,5.0,5.0,1.0,4.0,4.0,3.0,3.0,5.0,3.0,4.0,4.0,4.0,4.0,3.0,4.0,3.0,3.0,4.0
5.0,5.0,5.0,5.0,4.0,5.0,4.0,4.0,4.0,5.0,5.0,4.0,4.0,5.0,5.0,5.0,5.0,3.0,4.0,4.0
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
```

```
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
5.0,3.0,2.0,2.0,5.0,3.0,1.0,3.0,1.0,4.0,3.0,4.0,3.0,4.0,3.0,3.0,3.0,2.0,1.0,2.0
,
5.0,nan,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,5.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
3.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,3.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,4.0,3.0,4.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
4.0,5.0,nan,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,4.0,5.0,1.0,4.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,4.0,3.0,5.0
3.0,5.0,4.0,2.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0
5.0,2.0,1.0,3.0,3.0,4.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,1.0,3.0,1.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,1.0,2.0
5.0,5.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,3.0,2.0,3.0,4.0,2.0,1.0,2.0,2.0,1.0,2.0,2.0
5.0,2.0,4.0,3.0,4.0,2.0,2.0,2.0,2.0,4.0,3.0,3.0,3.0,5.0,2.0,2.0,2.0,4.0,2.0,2.0
5.0,5.0,5.0,5.0,4.0,3.0,3.0,3.0,3.0,5.0,3.0,4.0,4.0,3.0,2.0,2.0,3.0,3.0,3.0,3.0
5.0,5.0,4.0,3.0,5.0,4.0,4.0,4.0,4.0,5.0,4.0,4.0,5.0,4.0,3.0,3.0,4.0,3.0,3.0,4.0
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
1.0,4.0,1.0,4.0,3.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,4.0
4.0,2.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
2.0,5.0,3.0,2.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
5.0,5.0,5.0,5.0,3.0,3.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0
4.0,5.0,5.0,3.0,5.0,2.0,2.0,3.0,1.0,3.0,3.0,2.0,3.0,5.0,1.0,1.0,2.0,2.0,2.0,2.0
1.0,2.0,2.0,4.0,5.0,1.0,2.0,2.0,1.0,3.0,2.0,2.0,4.0,2.0,3.0,1.0,2.0,2.0,1.0,3.0
4.0,5.0,3.0,5.0,2.0,3.0,2.0,3.0,3.0,4.0,2.0,3.0,4.0,3.0,3.0,1.0,2.0,2.0,2.0,3.0
1.0,5.0,3.0,5.0,4.0,2.0,3.0,3.0,3.0,5.0,3.0,3.0,4.0,2.0,3.0,2.0,3.0,3.0,2.0,3.0
5.0,5.0,5.0,5.0,1.0,4.0,4.0,3.0,3.0,5.0,3.0,4.0,4.0,4.0,4.0,3.0,4.0,3.0,3.0,4.0
5.0,5.0,5.0,5.0,4.0,5.0,4.0,4.0,4.0,5.0,5.0,4.0,4.0,5.0,5.0,5.0,5.0,3.0,4.0,4.0
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
5.0,3.0,2.0,2.0,5.0,3.0,1.0,3.0,1.0,4.0,3.0,4.0,3.0,4.0,3.0,3.0,3.0,2.0,1.0,2.0
```

El código Python por el que se implementa el método se encuentra en el archivo *demo_bt500.py*.

demo_bt500.py:

```
import argparse
import csv
import sys
import pprint

import numpy as np
from scipy import linalg


def read_csv_into_3darray(csv_filepath):
    """
    Read data from CSV file.
```

```
    The data should be organized in a 2D matrix, separated by comma. Each row
    correspond to a PVS; each column corresponds to a subject. If a vote is
    missing, a 'nan' is put in place.

    If some subjects evaluated a PVS multiple times, another 2D matrix of the
    same size [num_PVS, num_subjects] can be added under the first one. A row
    with a single comma (,) should be placed before the repetition matrix.
    Where the repeated vote is not available, a 'nan' is put in place.

    :param csv_filepath: filepath to the CSV file.
    :return: the numpy array in 3D [num_PVS, num_subjects, num_repetitions].
    """

    data = []
    data3dlist = []
    with open(csv_filepath, 'rt') as datafile:
        datareader = csv.reader(datafile, delimiter=',')

        for row in datareader:
            if row != ["", ""]:
                data.append(np.array(row, dtype=np.float64))
            else:
                data3dlist.append(data)
                data = []
        data3dlist.append(data)

    data3d = np.zeros([len(data3dlist[0]), len(data3dlist[0][0]), len(data3dlist)])

    for r_idx, r_mat in enumerate(data3dlist):
        data3d[:, :, r_idx] = r_mat

    return data3d


def weighed_nanmean_2d(a, wts, axis):
    """
    Compute the weighted arithmetic mean along the specified axis, ignoring
    NaNs. It is similar to numpy's nanmean function, but with a weight.

    :param a: 1D array.
    :param wts: 1D array carrying the weights.
    :param axis: either 0 or 1, specifying the dimension along which the means
    are computed.
    :return: 1D array containing the mean values.
    """

    assert len(a.shape) == 2
    assert axis in [0, 1]
    d0, d1 = a.shape
    if axis == 0:
        return np.divide(
            np.nansum(np.multiply(a, np.tile(wts, (d1, 1)).T), axis=0),
            np.nansum(np.multiply(~np.isnan(a), np.tile(wts, (d1, 1)).T), axis=0)
        )
    elif axis == 1:
        return np.divide(
            np.nansum(np.multiply(a, np.tile(wts, (d0, 1))), axis=1),
            np.nansum(np.multiply(~np.isnan(a), np.tile(wts, (d0, 1))), axis=1),
        )
    else:
        assert False


def one_or_nan(x):
    """
    Construct a "mask" array with the same dimension as x, with element NaN
    where x has NaN at the same location; and element 1 otherwise.

    :param x: array_like
```

```
    :return: an array with the same dimension as x
    """
    y = np.ones(x.shape)
    y[np.isnan(x)] = float('nan')
    return y


def get_sos_j(sig_j, u_jkir):
    """
    Compute SOS (standard deviation of score) for presentation jk
    :param sig_j:
    :param u_jkir:
    :return: array containing the SOS for presentation jk
    """
    den = np.nansum(
        stack_3rd_dimension_along_axis(one_or_nan(u_jkir) / np.tile(sig_j ** 2,
(u_jkir.shape[1], 1)).T[:, :, None],
                                        axis=1),
        axis=1)
    s_jk_std = 1.0 / np.sqrt(np.maximum(0., den))
    return s_jk_std


def stack_3rd_dimension_along_axis(u_jkir, axis):
    """
        Take the 3D input matrix, slice it along the 3rd axis and stack the resulting 2D
matrices
        along the selected matrix while maintaining the correct order.
        :param u_jkir: 3D array of the shape [JK, I, R]
        :param axis: 0 or 1
        :return: 2D array containing the values
            - if axis=0, the new shape is [R*JK, I]
            - if axis = 1, the new shape is [JK, R*I]
    """

    assert len(u_jkir.shape) == 3
    JK, I, R = u_jkir.shape

    if axis == 0:
        u = np.zeros([R * JK, I])

        for r in range(R):
            u[r * JK:(r + 1) * JK, :] = u_jkir[:, :, r]

    elif axis == 1:
        u = np.zeros([JK, R * I])

        for r in range(R):
            u[:, r * I:(r + 1) * I] = u_jkir[:, :, r]

    else:
        NotImplementedError

    return u



def run_alternating_projection(u_jkir):
    """
    Run Alternating Projection (AP) algorithm.

    :param u_jkir: 3D numpy array containing raw votes. The first dimension
    corresponds to the presentation (jk); the second dimension corresponds to the
    subjects (i); the third dimension correspons to the repetitions (r).
    If a vote is missing, the element is NaN.

    :return: dictionary containing results keyed by 'mos_j', 'sos_j', 'bias_i'
    and 'inconsistency_i'.
    """
    JK, I, R = u_jkir.shape
```

```
    # video by video, estimate MOS by averaging over subjects
    u_jk = np.nanmean(stack_3rd_dimension_along_axis(u_jkir, axis=1), axis=1)  # mean
marginalized over i

    # subject by subject, estimate subject bias by comparing with MOS
    b_jir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None]
    b_i = np.nanmean(stack_3rd_dimension_along_axis(b_jir, axis=0), axis=0)  # mean
marginalized over j

    MAX_ITR = 1000
    DELTA_THR = 1e-8
    EPSILON = 1e-8

    itr = 0
    while True:

        u_jk_prev = u_jk

        # subject by subject, estimate subject inconsistency by averaging the
        # residue over stimuli
        e_jkir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None] - np.tile(b_i, (JK, 1))[:,
:, None]
        sig_i = np.nanstd(stack_3rd_dimension_along_axis(e_jkir, axis=0), axis=0)
        sig_j = np.nanstd(stack_3rd_dimension_along_axis(e_jkir, axis=1), axis=1)

        # video by video, estimate MOS by averaging over subjects, inversely
        # weighted by residue variance
        w_i = 1.0 / (sig_i ** 2 + EPSILON)
        # mean marginalized over i:
        u_jk = weighed_nanmean_2d(
            stack_3rd_dimension_along_axis(u_jkir - np.tile(b_i, (JK, 1))[:, :, None],
axis=1),
            wts=np.tile(w_i, R),  # same weights for the repeated observations
            axis=1)

        # subject by subject, estimate subject bias by comparing with MOS,
        # inversely weighted by residue variance
        b_jir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None]
        # mean marginalized over j:
        b_i = np.nanmean(stack_3rd_dimension_along_axis(b_jir, axis=0), axis=0)

        itr += 1

        delta_u_jk = linalg.norm(u_jk_prev - u_jk)

        msg = 'Iteration {itr:4d}: change {delta_u_jk}, u_jk {u_jk}, ' \
              'b_i {b_i}, sig_i {sig_i}'.format(
            itr=itr, delta_u_jk=delta_u_jk, u_jk=np.mean(u_jk),
            b_i=np.mean(b_i), sig_i=np.mean(sig_i))

        sys.stdout.write(msg + '\r')
        sys.stdout.flush()

        if delta_u_jk < DELTA_THR:
            break

        if itr >= MAX_ITR:
            break

    u_jk_std = get_sos_j(sig_j, u_jkir)
    sys.stdout.write("\n")

    mean_b_i = np.mean(b_i)
    b_i -= mean_b_i
    u_jk += mean_b_i

    return {
        'mos_j': list(u_jk),
        'sos_j': list(u_jk_std),
```

```
        'bias_i': list(b_i),
        'inconsistency_i': list(sig_i),
    }


if __name__ == "__main__":
    parser = argparse.ArgumentParser()

    parser.add_argument(
        "--input-csv", dest="input_csv", nargs=1, type=str,
        help="Filepath to input CSV file. The data should be organized in a 2D "
             "matrix, separated by comma. The rows correspond to PVSs; the "
             "columns correspond to subjects. If a vote is missing, input 'nan'"
             " instead.", required=True)

    args = parser.parse_args()
    input_csv = args.input_csv[0]

    o_jir = read_csv_into_3darray(input_csv)

    ret = run_alternating_projection(o_jir)

    pprint.pprint(ret)
```

Anexo 2
a la Parte 1

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 se ajusta a los métodos de evaluación descritos en las Partes 1 y 2 de la presente Recomendación.

Está constituido por ficheros de texto con la estructura que se muestra en los Cuadros 1-4 y 1-5. 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 1-4

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 de la pantalla = *entero*↵	Diagonal de la pantalla (pulgadas)
Marca y modelo de la pantalla = *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 1-5) incluyendo el trayecto
....	Nombre del fichero Resultados del cliente
Resultado(*j*).Nombre = *cadena de caracteres*↵	Identificación del laboratorio que efectúa la prueba
Resultado(*j*).Laboratorio = *cadena de caracteres*↵	Número total de observadores
Resultado(*j*).Número de observadores = *1 ≤ entero ≤ N*↵	Indica si los votos recogidos durante el entrenamiento se incluyen en el fichero DAT adjunto
Resultado(*j*).*Entrenamiento = «Sí» o «No»*↵	
[Resultado(*j*).Sesión (*i*).Observadores]↵	[Identificador de sección]
O(*k*).Nombre = *cadena de caracteres*↵	Identificación del observador
O(*k*).Apellido = *cadena de caracteres*↵	Opcional
O(*k*).Sexo = *«M» o «F»* ↵	Opcional
O(*k*).Edad = *entero*↵	Principales grupos socioeconómicos (por ejemplo, trabajador, estudiante)
O(*k*).Ocupación = *cadena de caracteres*↵	Distancia de observación en alturas de la pantalla (por ejemplo, 3 *H*, 4 *H*, 6 *H*)
O(*k*).Distancia = *entero*↵	

(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 1-5

Formato del fichero de texto de datos brutos Resultados.DAT

Formato y sintaxis del fichero nombre de fichero .DAT	Comentarios
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
entero entero entero.....↵	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 nota(1) del Cuadro 1-4.

Anexo 3
(informativo)
a la Parte 1

Característica de fallo de la imagen según su contenido

A3-1 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.

A3-2 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. 1-5.

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. 1-6 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. 1-7 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. 1-5.

FIGURA 1-5

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

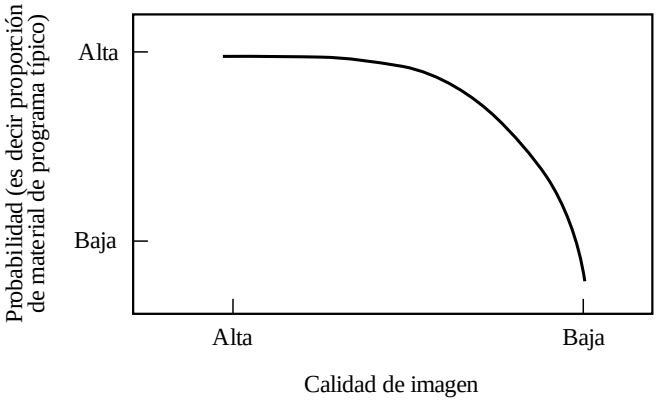


FIGURA 1-6

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

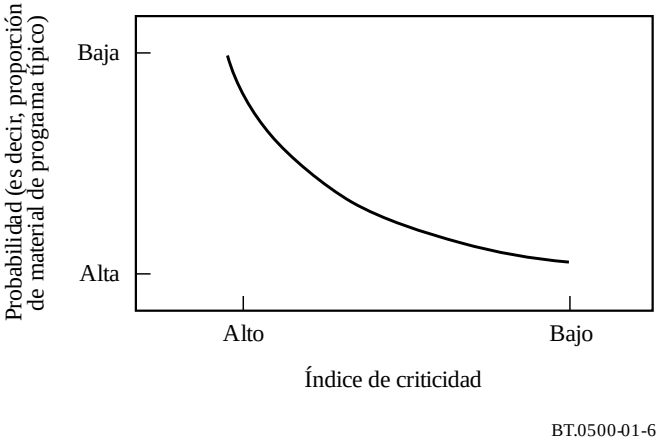
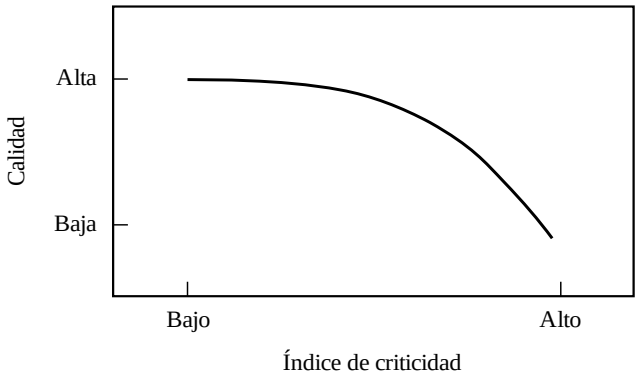


FIGURA 1-7

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



BT.0500-01-7

A3-3 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 alternativos (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 Parte 3 se da un ejemplo de característica de fallo de la imagen según su contenido para televisión digital.

Anexo 4
(informativo)
a la Parte 1

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

A4-1 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.

A4-2 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 Anexo 3). Esto se combinará con los resultados del análisis de las condiciones de transmisión para preparar la etapa final del proceso.

A4-3 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.

A4-4 Obtención de las características de fallo compuestas

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

El método básico utilizado es el procedimiento de doble estímulo con escala de calidad continua y, en particular, la versión recomendada de 10 s para secuencias en movimiento (véase el Anexo 2 a la Parte 2). Aquí, la referencia es una imagen con calidad de estudio en un formato apropiado (por ejemplo, un formato con resolución, frecuencia de trama, formato de imagen apropiado al sistema(s) en estudio). En contraste, la prueba presenta la misma imagen como si hubiera sido recibida por el sistema(s) en estudio bajo condiciones de canal seleccionado.

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

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

Anexo 5
(informativo)
a la Parte 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 6
(informativo)
a la Parte 1

Mediciones de información espacial y temporal

Las medidas de información espacial y temporal que figuran a continuación son de valor único para cada cuadro en una secuencia de prueba completa. Esto da lugar a una serie temporal de valores que por lo general tendrán un cierto grado de variación. Las medidas de información de percepción que figuran más adelante eliminan esta variabilidad con una función de máximo (máximo valor para la secuencia). La propia variabilidad se puede estudiar convenientemente, por ejemplo con muestras de información espacial-temporal cuadro a cuadro. La utilización de distribuciones de información a lo largo de una secuencia de prueba permite también una mejor evaluación de las escenas con cortes de escena

Información de percepción espacial (SI): Medida que generalmente indica el grado de detalle espacial de una imagen. Generalmente es mayor en escenas espacialmente más complejas. Esta información no constituye una medida de la entropía ni está asociada con la información definida en la teoría de la comunicación. La información de percepción espacial, SI, se basa en el filtro de Sobel. En primer lugar se filtra cada cuadro de vídeo (plano de luminancia) en un instante n (F_n) con el filtro de Sobel [Sobel(F_n)]. A continuación, se calcula la desviación típica de los píxeles ($std_{espacio}$) de cada cuadro filtrado con el filtro de Sobel. Esta operación se repite para cada cuadro de la secuencia de vídeo y da por resultado una serie temporal de información espacial de la escena. Se elige el máximo valor de la serie temporal (máx_{tiempo}) como representación del contenido de información espacial de la escena. Este proceso se puede representar en forma de ecuación como sigue:

$$\text{SI} = \text{máx}_{tiempo}\ \{std_{espacio}\ [\text{Sobel}(F_n)]\}$$

Información de percepción temporal (TI): Medida que generalmente indica la cantidad de cambios temporales de una secuencia de vídeo. Normalmente es mayor en secuencias de alta velocidad. Esta información no constituye una medida de la entropía ni está asociada con la información definida en la teoría de la comunicación.

La media de la información temporal, TI, se calcula como el máximo en el tiempo (máx_{tiempo}) de la desviación típica en el espacio ($std_{espacio}$) de $M_n(i, j)$ para todos i y j.

$$\text{TI} = \text{máx}_{tiempo}\ \{std_{espacio}\ [M_n(i, j)]\}$$

donde $M_n(i, j)$ es la diferencia entre píxeles en la misma posición en el cuadro, pero pertenecientes a dos cuadros consecutivos, es decir:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

donde $F_n(i, j)$ es el píxel de la i-ésima fila y j-ésima columna del n-ésimo cuadro en el tiempo.

NOTA – Para escenas que contienen cortes, pueden proporcionarse dos valores: uno en el que el corte de escena se incluye en la medición de información temporal y otro en el que se excluye de la medición.

Anexo 7
(informativo)
a la Parte 1

Términos y definiciones

Algoritmo	Una o varias técnicas de procesamiento de imagen
AVI	Audio vídeo intercalado (*audio video interleaved*)
CCD	Dispositivo de acoplamiento de cargas (*charge coupled device*)
CI	Intervalo de confianza (*confidence interval*)
CIF	Formato intermedio común (*common intermediate format*) (definido en la Recomendación UIT-T H.261 para videoteléfono: 352 líneas × 288 píxeles)
CRT	Tubo de rayos catódicos (*cathode ray tube*)
DSCQS	Método de escala de calidad continua de doble estímulo (*double stimulus using a continuous quality scale method*)
DSIS	Método de escala de degradación con doble estímulo (*double stimulus using an impairment scale method*)
Escena	Contenido audiovisual
LCD	Pantalla de cristal líquido (*liquid crystal display*)
MOS	Nota media de opinión (*mean opinion score*)
PDP	Panel de visualización de plasma (*plasma display panel*)
PS	Segmento de programa (*programme segment*)
QCIF	Cuarto de CIF (*quarter CIF*) (definido en la Recomendación H.261 para videoteléfono: 176 líneas × 144 píxeles)
SAMVIQ	Evaluación subjetiva de calidad de vídeo multimedios (*subjective assessment of multimedia video quality*)
SC	Método de comparación de estímulo (*stimulus comparison method*)
Secuencia	Escena con procesamiento combinado o sin procesamiento
SI	Información de percepción espacial (*spatial perceptual information*)
SIF	Formato intermedio normalizado (*standard intermediate format*) [definido en ISO 11172 (MPEG-1): 352 líneas × 288 píxeles × 25 cuadros/s y 352 líneas × 240 píxeles × 30 cuadros/s]
S/*N*	Relación señal/ruido (*signal-to-noise ratio*)
SP	Presentación simultánea (*simultaneous presentation*)
SS	Método de estímulo único (*single stimulus method*)
SSCQE	Método de evaluación de calidad continua de estímulo único (*single stimulus using a continuous quality evaluation method*)
std	Desviación típica (*standard deviation*)
TI	Información de percepción temporal (*temporal perceptual information*)

TP	Presentación de prueba (*test presentation*)
TS	Sesión de prueba (*test session*)
VTR	Magnetoscopio (*video tape recorder*)

PARTE 2

Descripción de las metodologías de evaluación subjetiva de las imágenes

Página

PARTE 2 45

Descripción de las metodologías de evaluación subjetiva de las imágenes 45

1 Introducción 47

2 Metodologías recomendadas para la evaluación de las imágenes 48

3 Observaciones 48

Anexo 1 a la Parte 2 – Método de escala de degradación con doble estímulo (DSIS) (método UER) ... 48

A1-1 Descripción general 48

A1-2 Disposición general 49

A1-3 Presentación del material de prueba 49

A1-4 Escalas de apreciación 50

A1-5 Introducción a las evaluaciones 50

A1-6 La sesión de prueba 51

Anexo 2 a la Parte 2 – Método de escala de calidad continua de doble estímulo (DSCQS) ... 51

A2-1 Descripción general 51

A2-2 Disposición general 51

A2-3 Presentación del material de prueba 52

A2-4 Escala de apreciación 53

A2-5 Análisis de los resultados 54

A2-6 Interpretación de los resultados 55

Anexo 3 a la Parte 2 – Métodos de estímulo único (SS) 55

A3-1 Disposición general 55

A3-2 Selección del material de prueba 55

A3-3 Sesión de prueba 56

A3-4 Tipos de métodos de estímulo único 56

A3-4.1 Métodos de apreciación por categorías de adjetivos 57

A3-4.2 Métodos de apreciación por categorías numéricas 57

A3-4.3 Métodos que no utilizan una escala de evaluación por categorías 57

Página

A3-4.4 Métodos de realización 58

Anexo 4 a la Parte 2 – Métodos de comparación de estímulos..... 58

A4-1 Disposición general 58

A4-2 Selección del material de prueba 58

A4-3 Sesión de prueba 58

A4-4 Tipos de métodos de comparación de estímulos 58

A4-4.1 Métodos de apreciación por categorías de adjetivos..... 59

A4-4.2 Métodos que no utilizan una escala de apreciación por categorías..... 59

A4-4.3 Métodos de realización 60

Anexo 5 a la Parte 2 – Evaluación de calidad continua de estímulo único (SSCQE) 60

A5-1 Dispositivo de registro y configuración..... 60

A5-2 Formato general del protocolo de prueba 60

A5-3 Parámetros de observación 61

A5-4 Escalas de apreciación 61

A5-5 Observadores 61

A5-6 Instrucciones a los observadores 61

A5-7 Presentación de datos y procesamiento y presentación de resultados 62

A5-8 Calibración de los resultados de calidad continuos y obtención de un único índice de calidad 62

Anexo 6 a la Parte 2 – Método de doble estímulo simultáneo para evaluación continua (SDSCE) 63

A6-1 Procedimiento de prueba 64

A6-2 Diferentes fases..... 64

A6-3 Características del protocolo de prueba..... 65

A6-4 Procesamiento de datos..... 65

A6-5 Fiabilidad de los sujetos..... 66

Anexo 7 a la Parte 2 – Evaluación subjetiva de la calidad de vídeo multimedios (SAMVIQ) 68

A7-1 Introducción..... 68

A7-2 Referencia explícita, referencia oculta y algoritmos 69

A7-3 Condiciones de la prueba..... 69

A7-4 Organización de la prueba 69

A7-5 Presentación y análisis de los datos 70

Página

A7-5.1 Información resumida 70

A7-5.2 Métodos de análisis 71

A7-5.3 Selección de los observadores 71

A7-6 Ejemplo de Interfaz para SAMVIQ (informativo) 72

Anexo 8 a la Parte 2 – Protocolo de observación para expertos (EVP) para la evaluación de la calidad de vídeo 72

A8-1 Montaje de laboratorio 72

A8-1.1 Selección y montaje de la pantalla 72

A8-1.2 Distancia de observación 73

A8-1.3 Condiciones de observación 73

A8-2 Observadores 73

A8-3 Célula básica de prueba 73

A8-4 Formulario de puntuación y escala de calificación 74

A8-5 Diseño de las pruebas y constitución de la sesión 75

A8-6 Capacitación 76

A8-7 Recopilación y tratamiento de los datos 76

A8-8 Condiciones de uso de los resultados del Protocolo de observación para expertos 76

A8-9 Limitaciones en el uso de los resultados del EVP 76

Adjunto 1 (informativo) al Anexo 8 a la Parte 2 – Aplicación del protocolo de observación por expertos y su comportamiento en presencia de un gran número de expertos evaluadores 77

1 Introducción

En esta Parte se detallan las distintas metodologías necesarias para llevar a cabo evaluaciones subjetivas de la calidad de las imágenes. En algunos casos, ello incluye una variación de las características de las evaluaciones comunes que figuran en el § 2 de la Parte 1.

Para garantizar que otros laboratorios puedan interpretar correctamente los resultados de las evaluaciones subjetivas de la calidad de las imágenes, es importante elaborar notas detalladas sobre los procedimientos y registrar cualquier variación de la metodología utilizada, incluida toda la información adicional que cualquier laboratorio que desee repetir el procedimiento de evaluación pueda requerir.

2 Metodologías recomendadas para la evaluación de las imágenes

Anexo 1 Método de escala de degradación con doble estímulo (DSIS)

Anexo 2 Método de escala de calidad continua de doble estímulo (DSCQS)

Anexo 3 Métodos de estímulo único (SS)

Anexo 4 Métodos de comparación de estímulos

Anexo 5 Evaluación de calidad continua de estímulo único (SSCQE)

Anexo 6 Método de doble estímulo simultáneo para evaluación continua (SDSCE)

Anexo 7 Evaluación subjetiva de la calidad de vídeo multimedios (SAMVIQ)

Anexo 8 Protocolo de observación para expertos (EVP) para la evaluación de la calidad de vídeo

3 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 de un planteamiento multidimensional.

Anexo 1
a la Parte 2

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

A1-1 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 Anexo 2 a la Parte 2).

El método de escala de degradación con doble estímulo (DSIS) (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.

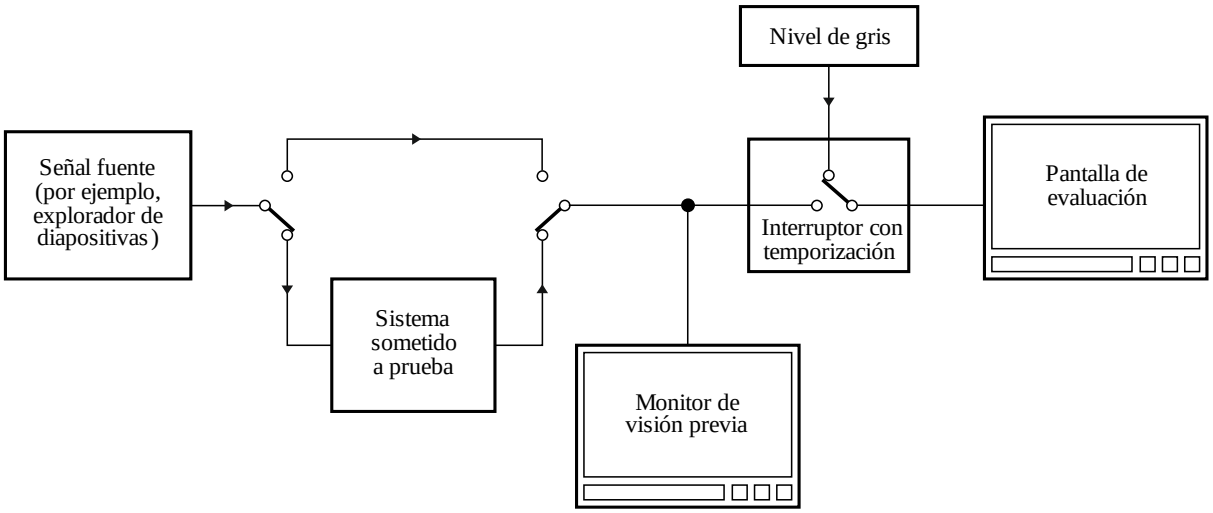
A1-2 Disposición general

En el § 2 de la Parte 1 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. 2-1.

FIGURA 2-1

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



BT.0500-02-1

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.

A1-3 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. 2-2a)

Variante II: La imagen o secuencia de referencia y la imagen o secuencia de prueba se presentan dos veces, como muestra la Fig. 2-2b).

Variante II: Esta variante requiere más tiempo que la variante I y puede aplicarse en los casos en que sea necesario diferenciar degradaciones muy pequeñas o se sometan a prueba secuencias en movimiento.

A1-4 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 sus notas.

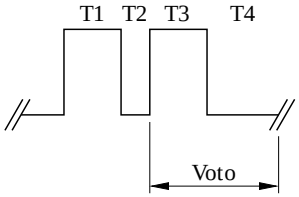
A1-5 Introducción a las evaluaciones

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.

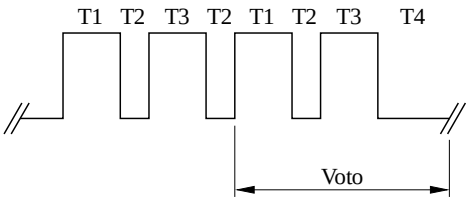
Debe pedirse a los observadores que observen la imagen durante los periodos T1 y T3. La votación debe autorizarse únicamente durante T4.

FIGURA 2-2

Estructura de presentación del material de prueba



a) Variante I



b) Variante II

BT.0500-02-2

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 de prueba
T4 =	5-11 s	Gris mediano

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

A1-6 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 Anexo 2 a la Parte 1 se presentan otras ideas sobre la selección de niveles de degradaciones.

Anexo 2
a la Parte 2

Método de escala de calidad continua de doble estímulo (DSCQS)

A2-1 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.

A2-2 Disposición general

En el § 2 de la Parte 1 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 § A1-6 del Anexo 1 a la Parte 2.

La disposición general del sistema de prueba debería ser la que se indica en la Fig. 2-3.

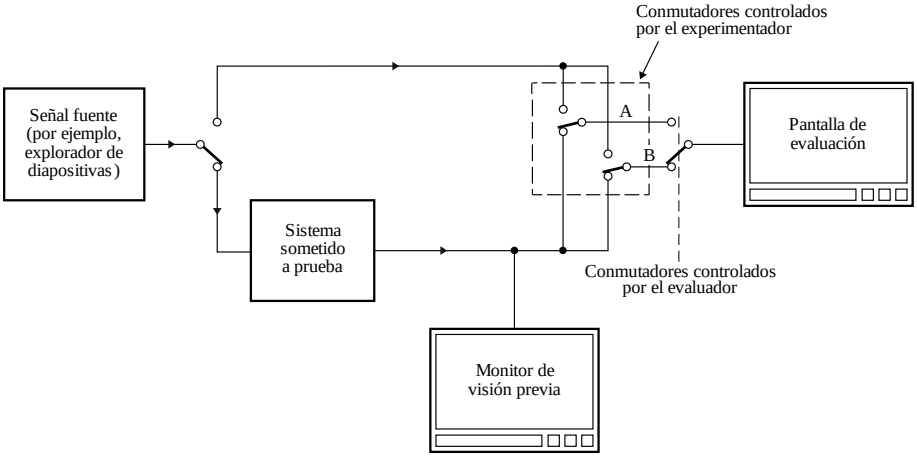
A2-3 Presentación del material de prueba

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. 2-4.

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 2-3

Disposición general del sistema de prueba para el método DSCQS



BT.0500-02-3

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

Variante I El evaluador, que suele estar solo, puede conmutar entre las dos condiciones A y B hasta que esté convencido de que se ha hecho una opinión de cada una. Las líneas A y B reciben la imagen directa de referencia, o la imagen transmitida por el sistema sometido a prueba, pero la transmisión por una la 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.

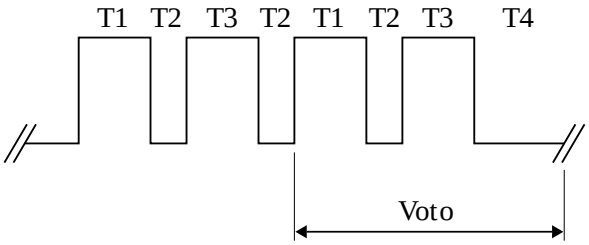
A2-4 Escala de apreciación

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. A2-5 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 2-4

Estructura de presentación del material de prueba



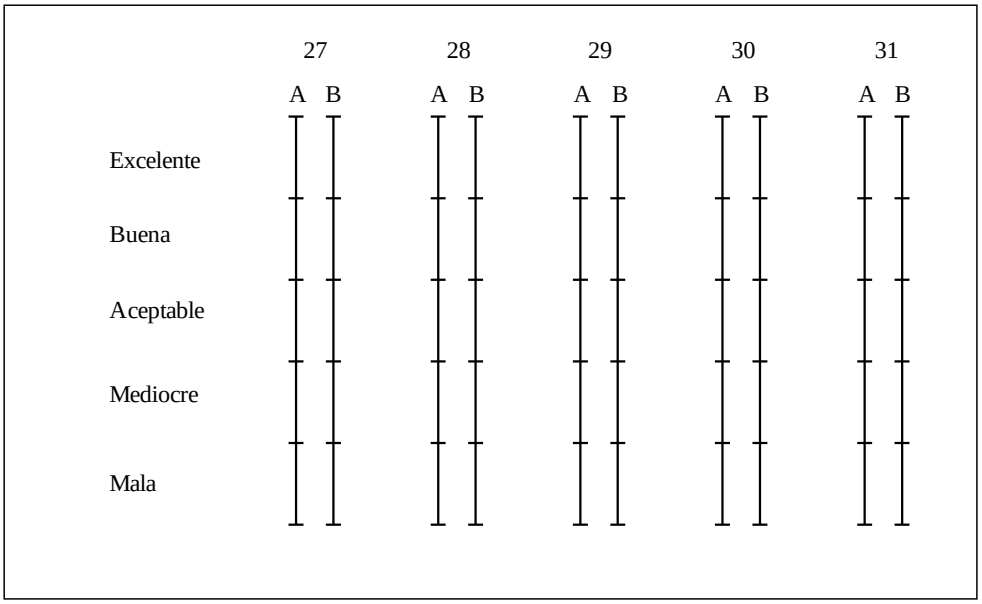
BT.0500-02-4

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

FIGURA 2-5

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



BT.0500-02-5

* 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.

A2-5 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 a la Parte 1 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 Anexo 2 a la Parte 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.

A2-6 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.

Anexo 3
a la Parte 2

Métodos de estímulo único (SS)

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.

A3-1 Disposición general

En el § 2 de la Parte 1 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.

A3-2 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 de la Parte 1.

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).

A3-3 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 de la Parte 1). 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 postexposición 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 postexposición.

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 la pantalla (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 postexposición 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 postexposición 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.

A3-4 Tipos de métodos de estímulo único

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

A3-4.1 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 2-1. 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 2-1

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.

A3-4.2 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.

A3-4.3 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).

A3-4.4 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.

Anexo 4
a la Parte 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.

A4-1 Disposición general

En el § 2 de la Parte 1 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.

A4-2 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.

A4-3 Sesión de prueba

En la prueba de evaluación se utilizará una pantalla, o bien dos pantallas debidamente sincronizadas, y se procederá en general como en los casos de estímulos únicos. Con una sola pantalla, 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 pantallas, 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.

A4-4 Tipos de métodos de comparación de estímulos

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

A4-4.1 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 2-2.

CUADRO 2-2

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.).

A4-4.2 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.

A4-4.3 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.

Anexo 5
a la Parte 2

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 técnica de evaluación de calidad continua de estímulo único (SSCQE).

A5-1 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.

A5-2 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 minutos;
– *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 minutos;

– *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.

A5-3 Parámetros de observación

Las condiciones de observación deberán ser las descritas actualmente en la Parte 1, o las condiciones específicas de la aplicación que figuran en la Parte 3.

A5-4 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 de la Parte 1.

A5-5 Observadores

Deberán participar al menos 15 observadores, no especializados, con las características que actualmente se recomiendan en el § 2.5 de la Parte 1.

A5-6 Instrucciones a los observadores

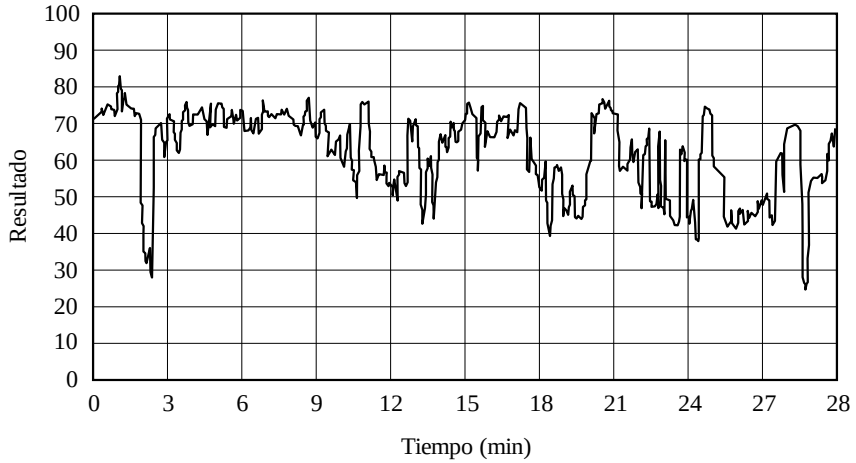
Si se evalúa la calidad de servicio (con audio de acompañamiento), deberá indicarse a los observadores que tengan en cuenta la calidad global, en vez de fijarse en la calidad vídeo solamente.

A5-7 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. 2-6).

FIGURA 2-6

Condición de prueba: códex X/segmento de programa Z



BT.0500-02-6

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. 2-7).

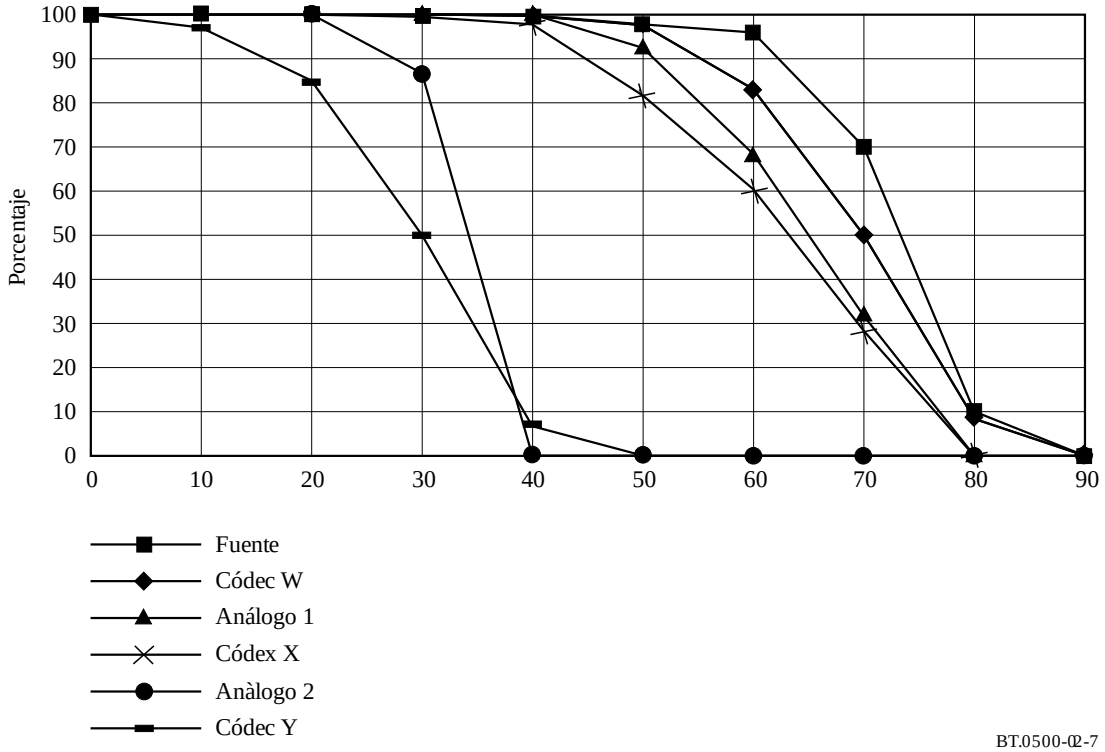
A5-8 Calibración de los resultados de calidad continuos y obtención de un único índice de calidad

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 2-7

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



Anexo 6
a la Parte 2

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 las anteriores metodologías normalizadas 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.

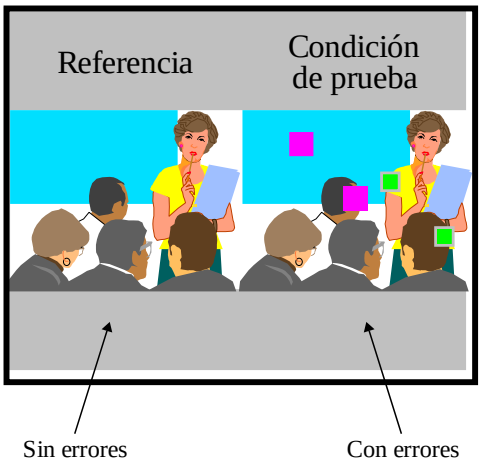
Como resultado, se ha elaborado y probado la siguiente nueva técnica SDSCE.

A6-1 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 la misma pantalla; en los demás casos se debe utilizar dos pantallas alineadas (véase la Fig. 2-8).

FIGURA 2-8

Ejemplo de formato de visualización



BT.0500-02-8

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.

A6-2 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.

A6-3 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.

A6-4 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. 2-9. Los resultados de este paso pueden ser representados en un diagrama temporal, como se muestra en la Fig. 2-10.

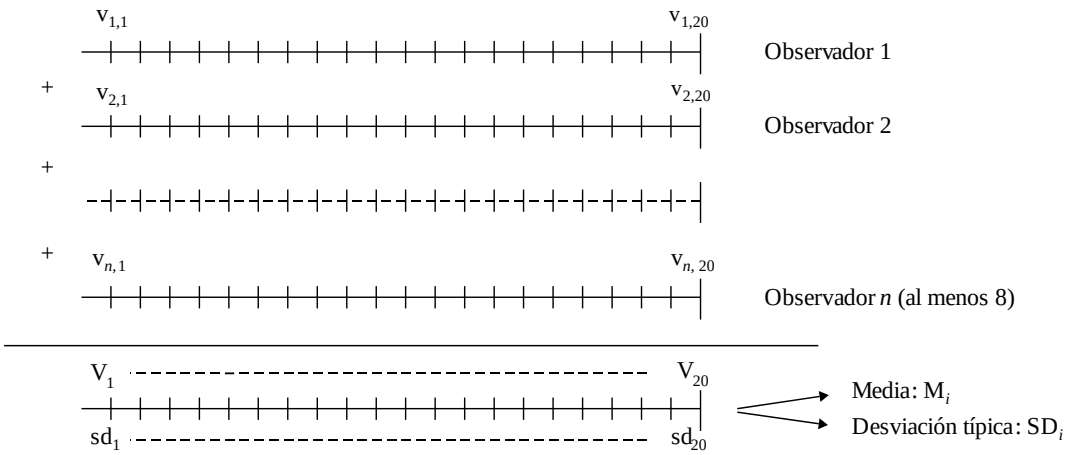
– Se analiza la distribución estadística de los valores medios calculados en el paso anterior (es decir, correspondiente a cada segmento de votos), y su frecuencia de aparición. Para evitar el efecto de novedad debido a las anteriores combinaciones de segmentos vídeo × condiciones de prueba, se rechazan 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. 2-11. 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.

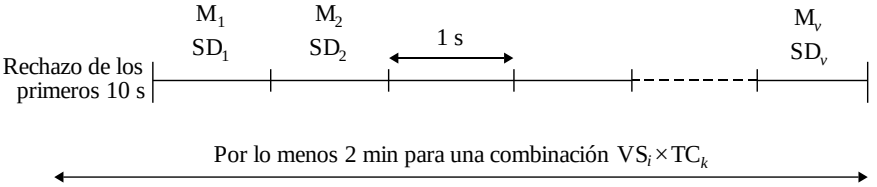
FIGURA 2-9

Procesamiento de datos

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



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



BT.0500-02-9

A6-5 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 § A1-2.3.2 del Anexo 1 a la Parte 1 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 2-10

Diagrama temporal

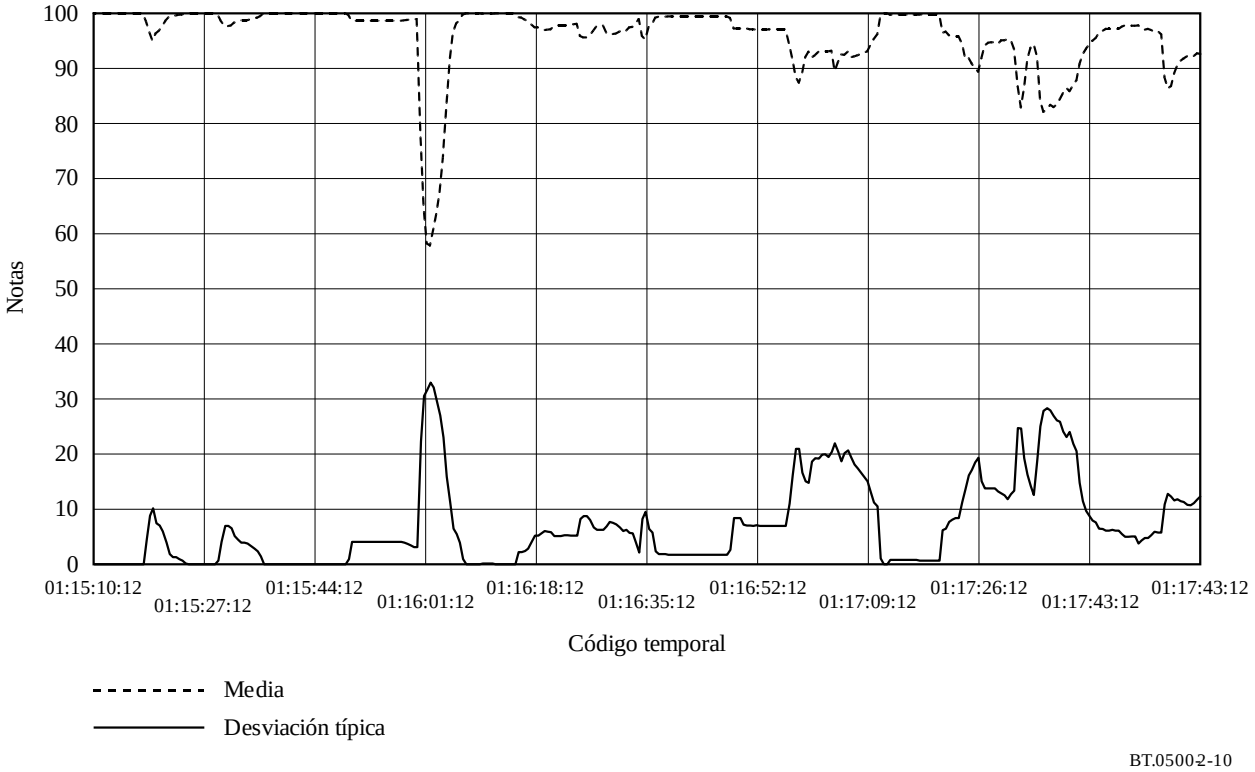
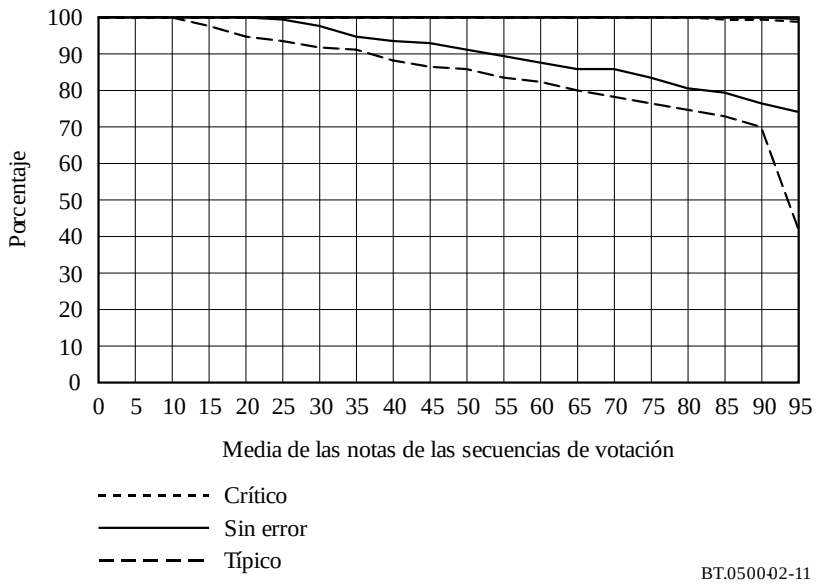


FIGURA 2-11

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



Anexo 7
a la Parte 2

Evaluación subjetiva de la calidad de vídeo multimedios (SAMVIQ)

A7-1 Introducción

El método de evaluación subjetiva de la calidad de vídeo multimedios (SAMVIQ) emplea una escala de calidad continua, que proporciona una medida de la calidad intrínseca de las secuencias de vídeo. Cada observador mueve un mecanismo deslizante en una escala graduada de 0 a 100, en la que están anotados 5 niveles de calidad dispuestos linealmente (excelente, bueno, normal, regular, malo).

En el marco del método SAMVIQ, al espectador se le da acceso a varias versiones de una secuencia. Cuando el espectador ha calificado todas las versiones, puede accederse al siguiente contenido de la secuencia.

El espectador selecciona aleatoriamente las diferentes versiones por medio de una interfaz de ordenador gráfica, pudiendo detener, revisar y modificar como desee la nota otorgada a cada versión de una secuencia. Este método incluye una secuencia de referencia explícita (es decir, no procesada) y varias versiones de la misma secuencia que incluyen tanto secuencias procesadas como no procesadas (es decir, referencia oculta). Cada versión de una secuencia se visualiza por separado, clasificándose usando una escala de calidad continua similar a la empleada en el método DSCQS. Así, el método se asemeja mucho funcionalmente a un método de estímulo único con acceso aleatorio, pero un observador puede ver la referencia explícita siempre que quiera, haciendo que este método se asemeje a uno que utilice una referencia.

El método de evaluación de calidad SAMVIQ usa una escala de calidad continua que proporciona una medida de la calidad intrínseca de las secuencias de vídeo. Cada observador mueve un mecanismo deslizante en una escala graduada de 0 a 100, en la que están anotados cinco niveles de calidad dispuestos linealmente (excelente, bueno, normal, regular, malo).

La evaluación de calidad se lleva a cabo sucesivamente para cada escena (véase la Fig. 2-12), incluyendo una *referencia explícita*, una *referencia oculta* y *varios algoritmos*.

Para entender mejor el método, se definen a continuación las siguientes palabras específicas:

Escena: contenido audiovisual.

Secuencia: escena con procesamiento combinado o sin procesamiento.

Algoritmo: una o varias técnicas de procesamiento de imagen.

A7-2 Referencia explícita, referencia oculta y algoritmos

Un método de evaluación incluye normalmente anclajes de calidad para estabilizar los resultados. En el método SAMVIQ se consideran dos anclajes de alta calidad por las siguientes razones. Se han llevado a cabo varias pruebas que indican desviaciones típicas minimizadas de las notas usando una *referencia explícita* en lugar de una oculta, o de no usar referencia. Concretamente, para evaluar el rendimiento de los códecs, es mejor utilizar una referencia explícita con objeto de lograr la máxima fiabilidad de los resultados. Se añade también una *referencia oculta* para evaluar la calidad intrínseca de la referencia, en lugar de la referencia explícita, porque la presentación es anónima, al igual que las secuencias procesadas. El nombre explícito «referencia» tiene influencia sobre aproximadamente el 30% de los observadores. Éstos dan la máxima nota posible (100) a la referencia explícita, siendo dicha nota totalmente diferente de la correspondiente a la referencia oculta. En particular, cuando no hay referencia disponible, la prueba sigue siendo posible, pero la desviación típica aumenta drásticamente.

El método SAMVIQ es adecuado para un contexto multimedios porque es posible combinar diferentes características de procesamiento de imagen, tales como tipo de códec, formato de imagen, velocidad binaria, actualización temporal, zooming, etc. La palabra *algoritmo* resume una de estas características, o combinación de las mismas.

A7-3 Condiciones de la prueba

La variación de la criticidad de una escena es limitada porque el contenido homogéneo se escoge en función de las mismas reglas empleadas implícitamente por otras metodologías que proporcionan una nota global (por ejemplo, los métodos de estímulo único). Una máxima duración de visualización de secuencia de 10 ó 15 s es, por tanto, suficiente para lograr una nota de calidad estabilizada y fiable. Los decodificadores-reproductores patentados, o una copia de pantalla de su salida, deben utilizarse para mantener la calidad de funcionamiento de la visualización adecuada.

A7-4 Organización de la prueba

a) La prueba se lleva a cabo sucesivamente para cada escena, como se describe en la Fig. 2-12.

b) Para la escena actual, es posible reproducir y otorgar una nota a cualquier secuencia en cualquier orden. Cada secuencia puede reproducirse y calificarse varias veces.

c) De una escena a otra, el acceso secuencial se aleatoriza, evitando que los observadores intenten votar de un modo idéntico con arreglo a un orden establecido. De hecho, en una prueba el orden del algoritmo permanece invariable para simplificar el análisis y la presentación de los resultados. Sólo se aleatoriza el acceso correspondiente a partir de un idéntico botón.

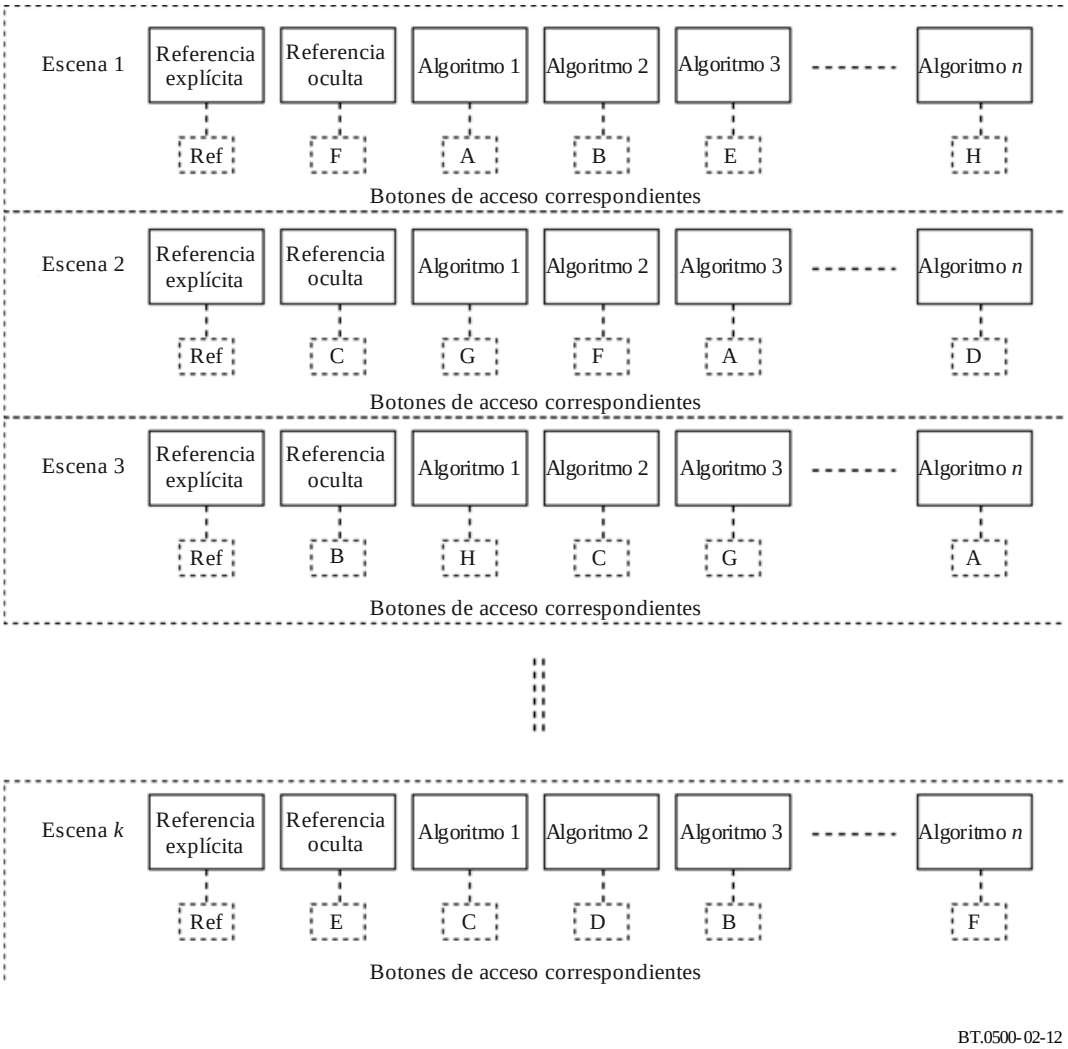
d) Para una primera visualización, la secuencia en curso debe reproducirse completamente antes de calificarse; de lo contrario, sería posible calificar y detenerse inmediatamente.

e) Para probar la siguiente escena deben calificarse todas las secuencias de la escena en curso.

f) Para concluir la prueba, deben calificarse todas las secuencias de cada escena.

FIGURA 2-12

Ejemplo de organización de prueba para el método SAMVIQ



El método SAMVIQ se implementa mediante programas informáticos. Además de los botones de acceso mostrados en la Fig. 2-12, se precisan los botones «reproducir», «detener», «escena siguiente» y «escena anterior», para permitir al espectador gestionar la presentación de las diferentes escenas (véase el § A7-6 como ejemplo). Cuando un espectador da una nota, ésta debe aparecer bajo el botón de acceso correspondiente a dicha escena. Cuando todas las versiones diferentes de una secuencia han sido calificadas, el espectador aún puede comparar las notas y modificar, si es necesario, sus valores. No es preciso revisar completamente la secuencia en curso porque ya han quedado patentes diferencias importantes durante la primera visualización.

A7-5 Presentación y análisis de los datos

A7-5.1 Información resumida

Se necesita información precisa sobre el entorno de la prueba para reproducir la misma, o para comparar los resultados de diferentes pruebas. Por tanto, se sugiere comunicar la información sobre el entorno de prueba como se describe en el Cuadro 2-3.

CUADRO 2-3

Información resumida de las pruebas

Nombre del método	
Tecnología de visualización	
Nombre de referencia de la visualización	
Valor de cresta de la luminancia (cd/m²)	
Nivel de luminancia del negro (cd/m²)	
Configuración del nivel de negro: PLUGE (umbral percibido de la distancia del nivel de negro a ultra negro = 8). De lo contrario, indica el valor umbral	
Nivel de luminancia de fondo (cd/m²)	
Iluminación (lux)	
Distancia de observación: – No limitada: parte frontal de la visualización – Limitada: nH	
Tamaño de la visualización (diagonal en pulgadas)	
Relación de visualización entre anchura y altura	
Formato de visualización (N° de columnas, N° de líneas)	
Formato de entrada de imagen (N° de columnas, N° de líneas)	
Formato de salida de imagen(1) (N° de columnas, N° de líneas)	
Temperatura de color blanco: D65 de lo contrario Coordenadas de color blanco (*x*, *y*)	
Número de observadores efectivos	

(1) Esta información se necesita cuando se procesa la imagen de entrada, por ejemplo al reescalarla, tras la visualización.

Las características de visualización pueden influir en los resultados de la prueba. Para monitores de pantalla plana se debe comunicar información adicional tal como la respuesta de luminancia (fidelidad gamma) y los colores primarios.

Las características de las secuencias de vídeo son importantes para diseñar una prueba o explicar sus resultados. Se sugiere comunicar las características espaciales y temporales como se describe en el Anexo 1 a la Parte 1. Debe considerarse esta información en la colección de secuencias de prueba en la biblioteca de material vídeo adecuado para la evaluación subjetiva de la calidad de vídeo en aplicaciones multimedios.

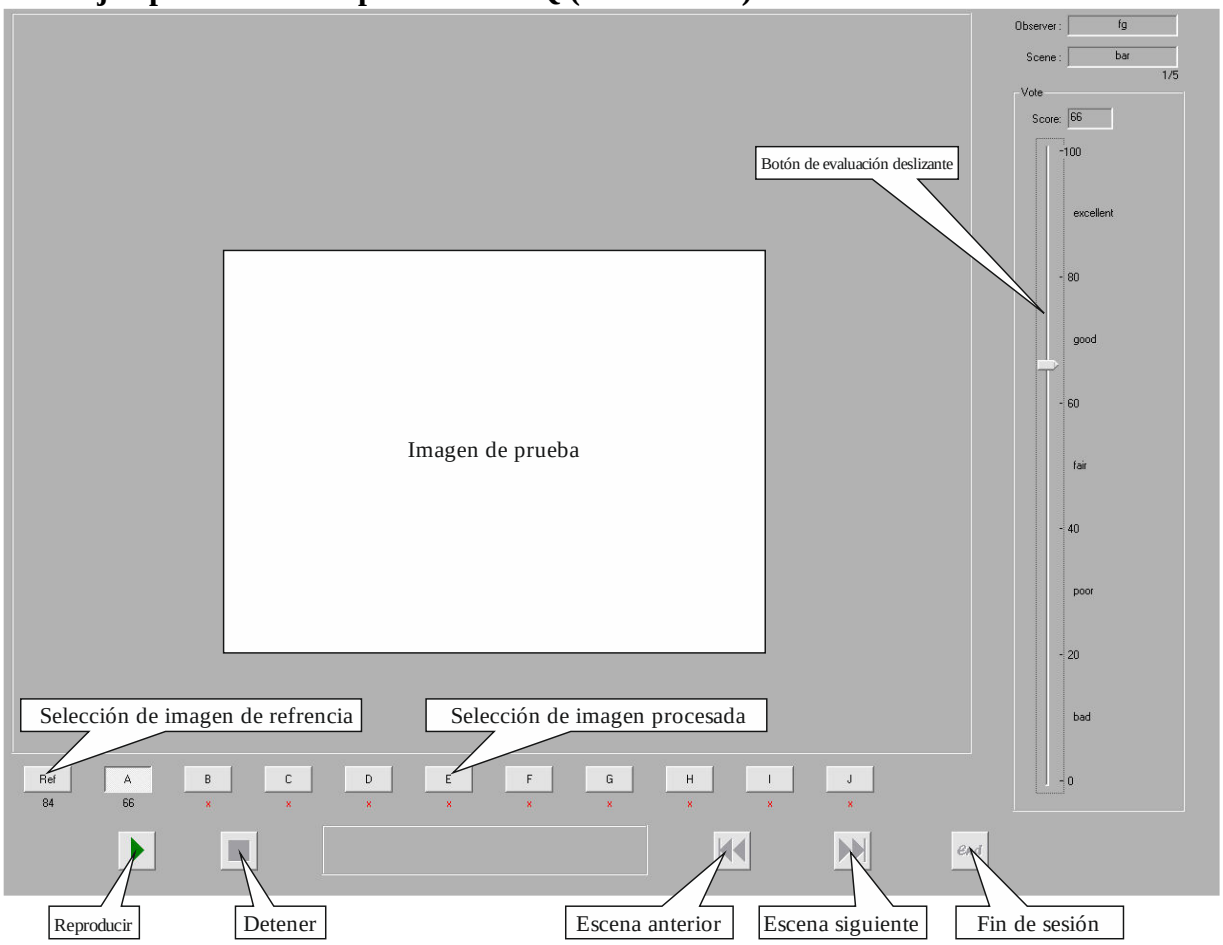
A7-5.2 Métodos de análisis

Los métodos de análisis son los descritos en el Anexo 1 a la Parte 1.

A7-5.3 Selección de los observadores

El análisis para SAMVIQ se describe en el § A1-2.3.4 del Anexo 1 a la Parte 1.

A7-6 Ejemplo de Interfaz para SAMVIQ (informativo)



BT.0500- 02-12a

Anexo 8
a la Parte 2

Protocolo de observación para expertos (EVP) para la evaluación de la calidad de vídeo

En este Anexo se describe el método para la evaluación subjetiva de la calidad de vídeo de imágenes en movimiento aplicando el protocolo de observación para expertos, con la participación de un reducido número de espectadores, todos ellos seleccionados entre expertos en el ámbito del procesamiento de vídeo de que se trate.

A8-1 Montaje de laboratorio

A8-1.1 Selección y montaje de la pantalla

La pantalla debe de ser una pantalla plana con características para aplicaciones profesionales (por ejemplo, estudios o unidades móviles de radiodifusión); la dimensión diagonal de la pantalla puede variar entre 22 pulgadas (mínimo) y 40 pulgadas (recomendado), aunque puede llegar a 50 pulgadas o más cuando se evalúan sistemas de imagen con una resolución de televisión de alta definición o superior.

Se permite utilizar una porción reducida de la zona activa de visualización en la pantalla, en cuyo caso la zona en torno a la parte activa de la pantalla se debe ajustar a gris medio. En estas condiciones de uso no se debe permitir que la pantalla tenga una resolución diferente de la del original.

La pantalla debe permitir el ajuste y la calibración adecuada de la luminancia y el color mediante un fotómetro profesional. Para llevar a cabo la prueba la calibración, la pantalla debe cumplir los parámetros especificados en la Recomendación pertinente.

A8-1.2 Distancia de observación

La distancia de observación a la que se situarán los expertos se debe elegir de conformidad con la resolución de la pantalla, a la altura de la parte activa de la pantalla, en función de la distancia de observación nominal que se describe en el § 2.1.3.2 de la Parte 1, aunque puede ser menor con arreglo a los requisitos relativos a las condiciones críticas de observación.

A8-1.3 Condiciones de observación

No es imprescindible llevar a cabo los ensayos de protocolos de observación para expertos (EVP) en un laboratorio de pruebas, pero si es importante proteger el emplazamiento de las pruebas de perturbaciones acústicas o visuales (por ejemplo, se puede utilizar también un despacho o una sala de reuniones tranquilos).

Se debe eliminar cualquier fuente de luz directa o indirecta que incida en la pantalla; la luz ambiente debe ser tenue y se mantendrá con la menor intensidad posible que permita rellenar los formularios (si procede).

El número de expertos sentados frente a la pantalla puede ser variable en función del tamaño de la pantalla con el fin de garantizar una idéntica reproducción de la imagen y la misma presentación de estímulos para todos los observadores.

A8-2 Observadores

Los observadores que participen en un experimento EVP deben ser expertos en la materia de estudio.

Los observadores no se seleccionarán necesariamente por su agudeza visual o daltonismo puesto que debe tratarse de personas cualificadas.

El número mínimo de observadores diferentes será de nueve.

Para lograr el número mínimo de observadores se puede repetir la prueba realizando el mismo experimento en la misma ubicación o en más de una ubicación. Las calificaciones obtenidas en diferentes ubicaciones que participen en una sesión de observación para expertos se pueden procesar juntas de forma estadística.

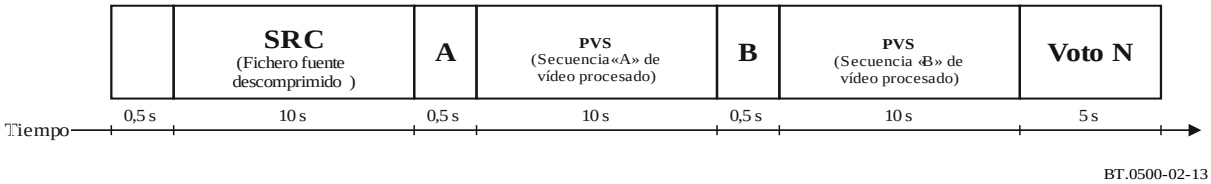
A8-3 Célula básica de prueba

El material que se presente a los expertos se debe organizar a partir de una célula básica de prueba (BTC) para cada pareja de condiciones de codificación sometida a evaluación (véase la Fig. 2-13).

Los videoclips de secuencias fuente de referencia (SRC) y de secuencias de vídeo procesadas (PVS) que se consideren en una BTC siempre se referirán a la misma secuencia de vídeo con el fin de que los expertos puedan identificar cualesquiera mejoras en la calidad visual aportadas por los algoritmos de compresión sometidos a prueba.

FIGURA 2-13

Temporización de una célula básica de prueba para el Protocolo de observación para expertos



La BTC se organizará de la forma siguiente:

– 0,5 segundos con la pantalla ajustada a gris medio (valor medio de la escala de luminancia);
– 10 segundos de presentación del videoclip de referencia descomprimido;
– 0,5 segundos para mostrar el mensaje «A» (primer vídeo sometido a evaluación) sobre un fondo gris medio;
– 10 segundos de presentación de una versión deficiente del videoclip;
– 0,5 segundos para mostrar el mensaje «B» (segundo vídeo sometido a evaluación) sobre un fondo gris medio;
– 10 segundos de presentación de una versión deficiente del videoclip;
– 5 segundos para mostrar un mensaje que solicite la opinión de los observadores.

Al mensaje «Voto» debe seguir un número que facilite la sincronización del formulario de puntuación.

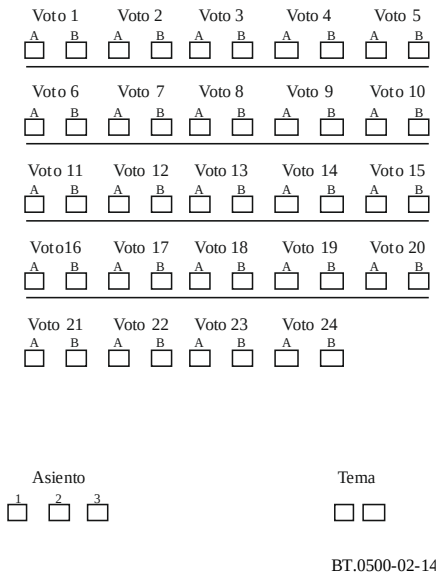
A8-4 Formulario de puntuación y escala de calificación

Como se muestra en la Fig. 2-13, la presentación de los videoclips se debe realizar de forma que se muestre en primer lugar la referencia sin deficiencias (SRC), seguida de dos secuencias de vídeo con deficiencias (PVS). El orden de presentación de la PVS se modificará de forma aleatoria para cada BTC y los observadores no deben conocer el orden de la presentación.

FIGURA 2-14

Ejemplo de formulario de puntuación para una sesión de observación para expertos de 24 BTC

Sesión 1



Se utiliza una escala numérica de 11 notas de 10 (degradación imperceptible) a 0 (degradación muy molesta).

El Cuadro 2-4 indica el significado de las 11 notas de la escala numérica.

CUADRO 2-4

Significado de las 11 notas de la escala numérica

Nota	Elemento de degradación	
10	Imperceptible	
9	Ligeramente perceptible	En parte
8		En todo
7	Perceptible	En parte
6		En todo
5	Claramente perceptible	En parte
4		En todo
3	Molesto	En parte
2		En todo
1	Muy molesto	En parte
0		En todo

Se solicita a los observadores que rellenen un cuestionario con dos casillas (denominadas «A» y «B») para cada BTC para que anoten en cada una de ellas la calificación seleccionada de la escala numérica de 11 notas.

La Fig. 2-14 muestra un ejemplo de formulario de puntuación para una sesión constituida por 24 BTC.

Para cada BTC, los observadores rellenan tanto la casilla identificada por la letra **A** (para evaluar el videoclip presentado en primer lugar) como la identificada con la letra **B** (para evaluar el videoclip presentado en segundo lugar).

La presentación del videoclip original sin degradaciones permite a los expertos evaluar cualquier degradación con mayor facilidad.

Durante las «sesiones de capacitación» se debe explicar detalladamente el significado de la escala numérica de 11 notas como se describe a continuación.

A8-5 Diseño de las pruebas y constitución de la sesión

El diseñador de la prueba establecerá aleatoriamente el orden de presentación de las BTC, de forma que no se muestre el mismo videoclip con o sin degradación dos veces consecutivas.

Todas las sesiones de observación deberán comenzar con una «fase de estabilización» que incluya las BTC «mejor» y «peor» así como dos sesiones de observación de «calidad media» elegidas entre las que figuran en cada sesión de prueba. Esto permite a los observadores formarse una opinión inmediata sobre la gama de calidades desde el inicio de la sesión de pruebas.

Si la sesión de observación dura más de 20 minutos, el diseñador de la prueba deberá dividirla en dos (o más) sesiones de observación diferenciadas que no excedan de 20 minutos cada una. En este caso, se incluirá la «fase de estabilización» antes de cada sesión de observación.

A8-6 Capacitación

Aunque este procedimiento esté previsto para participantes expertos, antes de cada experimento se recomienda organizar una sesión corta de capacitación (con 5 ó 6 BTC).

Los vídeos utilizados durante la sesión de capacitación pueden ser los mismos que los que se utilicen durante la sesión real, aunque el orden de presentación debe ser diferente.

Los observadores deben ejercitarse en el uso de la escala de 11 notas analizando cuidadosamente los videoclips que se muestren inmediatamente después de los mensajes «A» y «B» en la pantalla y deben comprobar si pueden advertir alguna diferencia con el videoclip mostrado en primer lugar (la SRC).

A8-7 Recopilación y tratamiento de los datos

Las notas se deben recopilar al final de cada sesión e introducir en una hoja de cálculo electrónica para obtener los valores MEDIOS.

Se aconseja realizar una «selección posterior» de los observadores utilizando una correlación lineal de Pearson.

La función «correlación» se aplicará tomando en consideración todas las notas de cada sujeto en relación con las notas medias de opinión (MOS); se puede fijar un valor umbral para determinar cada observador como «aceptable» o «rechazado» (la Recomendación UIT-T P.910 sugiere que los sujetos se rechacen por debajo de un valor umbral igual a 0,75).

A8-8 Condiciones de uso de los resultados del Protocolo de observación para expertos

El Protocolo de observación para expertos (EVP) se puede emplear cuando no se pueda llevar a cabo un experimento de evaluación subjetiva formal debido a falta de tiempo o de recursos.

El EVP requiere menos tiempo que la evaluación subjetiva formal y puede ejecutarse en un entorno «informal», en el supuesto de que el entorno en el que se lleva a cabo esté protegido de perturbaciones externas visuales y auditivas.

Las únicas condiciones obligatorias se refieren a las condiciones ambientales de iluminación y de observación (pantalla, ángulo de observación y distancia de observación) que se describen en los párrafos anteriores.

A8-9 Limitaciones en el uso de los resultados del EVP

Aunque el EVP está proporcionando resultados aceptables con solo nueve observadores, las MOS obtenidas por un experimento EVP no pueden sustituir los resultados obtenidos mediante un experimento de evaluación subjetiva formal.

Los datos de las MOS obtenidos mediante el EVP se podrían utilizar para disponer de una indicación preliminar del grado de perturbación.

Los datos de las MOS obtenidos mediante el EVP se podrían utilizar para efectuar una clasificación preliminar de los esquemas de procesamiento de vídeo sometidos a evaluación.

Cuando se considere conveniente o necesario, se puede llevar a cabo un experimento EVP en paralelo en diferentes ubicaciones, siempre que las condiciones de observación, la distancia de observación y el diseño de la prueba sean idénticos.

Si el experimento se realiza en diferentes ubicaciones y el número de observadores expertos implicados en el mismo EVP es superior a 15, se podrían procesar los datos subjetivos brutos para obtener los datos de MOS, desviación típica e intervalo de confianza que pueden contribuir a establecer una clasificación más precisa de los casos sometidos a prueba. En este último caso se pueden realizar análisis de estadística inferencial más precisos, por ejemplo, pruebas T-Student.

Adjunto 1
(informativo)
al Anexo 8
a la Parte 2

Aplicación del protocolo de observación por expertos y su comportamiento en presencia de un gran número de expertos evaluadores

En este Adjunto informativo se presentan los resultados de dos evaluaciones subjetivas de vídeos con codificación HD y UAD utilizando EVP, realizadas durante la 117ª reunión del MPEG. Para realizar las evaluaciones se aplicaron las disposiciones del Anexo 8 a fin de establecer una clasificación rápida y fiable de dos métodos de codificación fuente distintos.

Dado que a la 117ª reunión del MPEG asistió un elevado número de expertos, el número de evaluadores que participaron en las dos sesiones EVP superó con mucho los 9 que se aconsejan en el Anexo 8 a la Parte 2 de la presente Recomendación. A la prueba EVP HD asistieron 30 expertos y a la prueba EVP UAD asistieron 32 expertos.

La elevada participación de evaluadores expertos brindó la oportunidad de analizar los datos MOS a fin de verificar el nivel de fiabilidad inherente al uso del Anexo 8 para clasificar vídeos codificados.

Para esta evaluación se formaron cuatro grupos de observadores (a saber, de 9, 12, 15 y 18 observadores) y se compararon los valores MOS obtenidos por el grupo de 9 expertos con los obtenidos por los grupos de 12, 15 y 18 expertos observadores.

El objetivo era comparar la clasificación establecida por los 9 expertos (y, por tanto, conforme con el protocolo EVP) y las clasificaciones establecidas por los grupos de 12, 15 y 18 expertos (semejante a evaluación subjetiva formal).

Como puede verse en la Fig. 2-15 (experimento con contenido UAD) y en la Fig. 2-16 (experimento con contenido HD), los resultados de las clasificaciones son muy semejantes en los cuatro casos considerados.

Tomando los resultados obtenidos por el grupo de 18 observadores como una especie de «verdad fundamental», se pueden trazar los gráficos de las Figs. 2-15 y 2-16 clasificando los puntos de prueba según los valores MOS obtenidos por el grupo de 18 observadores (línea roja continua).

Las demás líneas del gráfico muestran los resultados obtenidos por los grupos de 9 (línea roja discontinua), 12 (línea azul discontinua) y 15 (línea verde continua) observadores.

De la comparación de los resultados representados en las Figs. 2-15 y 2-16 se desprende lo siguiente:

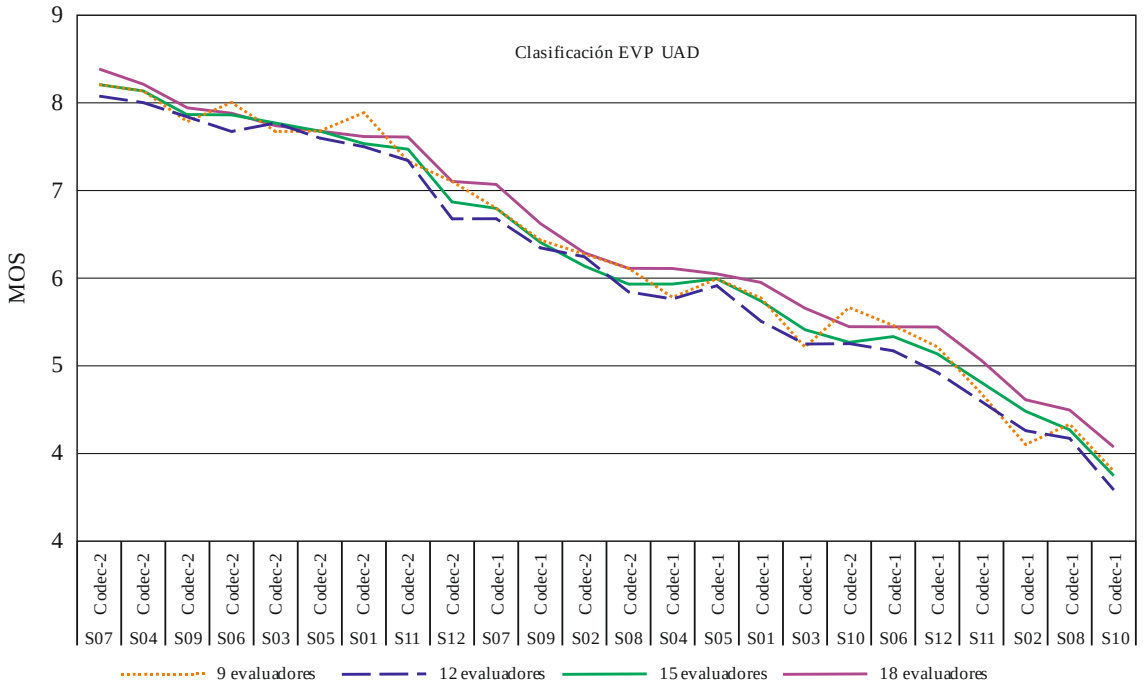
– los gráficos de los grupos de 15 y 18 observadores muestran una pendiente homogénea de los valores MOS de mayor calidad a menor calidad;

– los gráficos de los grupos de 9 y 12 observadores muestran algunas «inversiones» de la clasificación con respecto al grupo de 18 observadores, aunque la variación de la puntuación es bastante limitada.

Como conclusión puede decirse que los experimentos EVP aquí descritos demuestran un muy buen rendimiento del protocolo EVP, lo que confirma lo indicado en el Anexo 8, a saber, que la utilización del protocolo EVP, aunque no puede considerarse un sustituto completo del experimento subjetivo formal, puede considerarse un procedimiento de evaluación estable con el que se obtienen resultados muy cercanos a los que se consiguen cuando se dispone de muchos más observadores y se realiza una evaluación subjetiva formal.

FIGURA 2-15

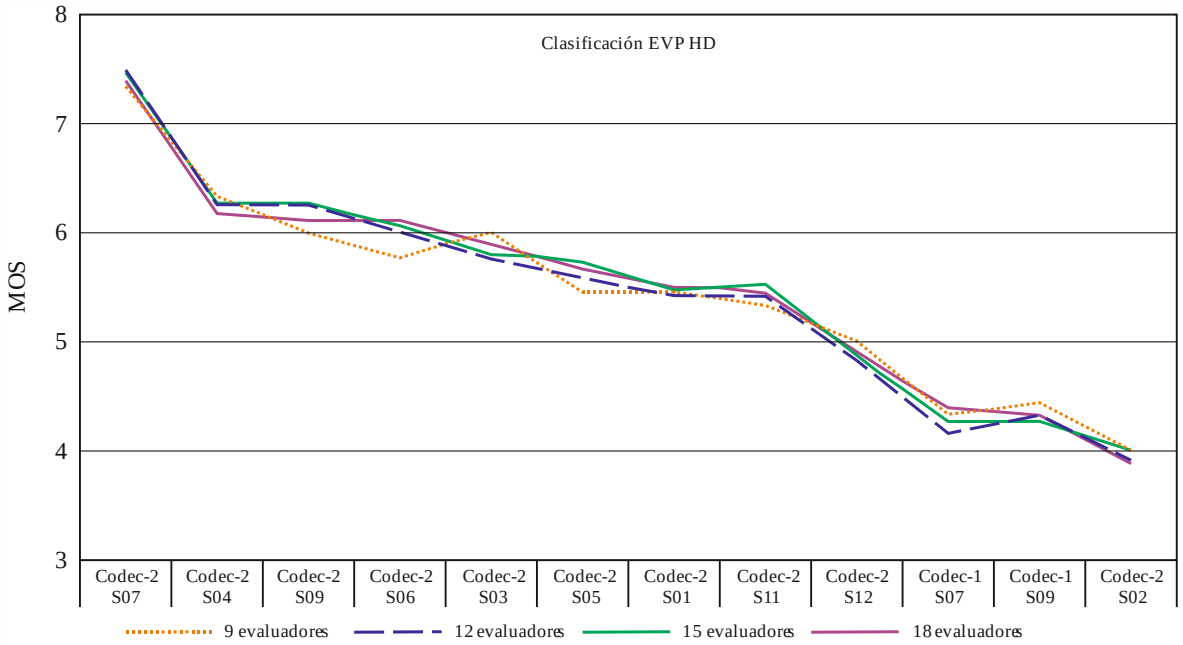
Clasificación del experimento UAD en función del número de evaluadores



BT.0500-02-15

FIGURA 2-16

Clasificación del experimento HD en función del número de evaluadores



BT.0500-02-16

PARTE 3

Metodologías para la evaluación subjetiva de la calidad de las imágenes específicas de aplicaciones

A fin de concebir una prueba de evaluación subjetiva, cabría tener en cuenta diversos aspectos específicos de la aplicación. En esta Parte 3, se facilitan orientaciones para la evaluación subjetiva de la calidad de las imágenes en determinados formatos de imagen y aplicaciones:

Anexo 1 Evaluación subjetiva de sistemas de televisión de definición estándar (TVDS)

Anexo 2 Métodos de evaluación subjetiva de la calidad de las imágenes en sistemas de televisión de alta definición

Anexo 3 Evaluación subjetiva de la calidad de las imágenes alfanuméricas y gráficas en servicios de teletexto y otros servicios de texto similares

Anexo 4 Evaluación subjetiva de la calidad de las imágenes de los servicios multiprograma

Anexo 5 Observación especializada de la calidad de las imágenes de los sistemas destinados a la proyección digital de imágenes digitales en pantalla grande en cines

Anexo 6 Metodología para la evaluación subjetiva de la calidad de vídeo en aplicaciones multimedios

Anexo 7 Métodos de evaluación subjetiva de los sistemas de televisión 3D estereoscópica

Anexo 1
a la Parte 3

Evaluación subjetiva de sistemas de televisión de definición estándar (TVDS)

A1-1 Introducción

En este Anexo, que se utilizará conjuntamente con las Partes 1 y 2 de la presente Recomendación, se proporcionan detalles sobre la aplicación de los métodos generales indicados en la Recomendación a las evaluaciones subjetivas de los sistemas digitales cuyo nivel de calidad es el de los sistemas de televisión convencional o aproximado; los detalles de procedimiento que se ofrece en esta Recomendación, junto con la información general, tratan de las pruebas de códecs (o sistemas) utilizados para cursar material elaborado según la Recomendación UIT-R BT.601 por aplicaciones de contribución y distribución, así como los utilizados en las aplicaciones de emisión.

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

destinados a aplicaciones de contribución significa que los procedimientos de medición que se dan en la presente Recomendación pueden también utilizarse para su evaluación.

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

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

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

A1-2 Condiciones de observación

Las condiciones de observación generales para las evaluaciones subjetivas son las que se indican en el § 2 de la Parte 1. Las condiciones de observación específicas para la evaluación subjetiva de los sistemas digitales se indican en los apartados que figuran a continuación.

A1-2.1 Entorno del laboratorio

El entorno del laboratorio crea las condiciones críticas para la prueba de sistemas. Las condiciones de observación específicas para la evaluación subjetiva en laboratorio se indican en el Cuadro 3-1.

CUADRO 3-1

Condiciones de observación específicas para la evaluación subjetiva de sistemas digitales en laboratorio

Condición	Descripción	Valores
a	Relación entre la distancia de observación y la altura de la imagen	4 *H* y 6 *H* (1)
b	Luminancia de cresta	70 cd/m^2
c	Ángulo de observación subtendido por la parte del fondo que satisface las especificaciones	≥ 43° de altura × 57° de anchura
d	Imagen	Pantalla de alta calidad. Tamaño ≥ 20 pulgadas (50 cm)(2)

(1) 6 *H* es la distancia de observación nominal (DVD, design viewing distance) para evaluar sistemas digitales de definición estándar, pero también es aceptable utilizar evaluadores con 4 *H*, siempre y cuando los resultados se indiquen por separado.

(2) Como hay ciertas indicaciones de que el tamaño de la imagen puede influir en los resultados de la evaluación subjetiva, se pide a quienes realizan los experimentos que comuniquen el tamaño de la pantalla y la marca y el modelo del aparato utilizado en los mismos.

A1-2.2 Entorno doméstico

Este entorno ofrece los medios necesarios para evaluar la calidad de la cadena de televisión digital desde el punto de vista del consumidor. Las condiciones de observación específicas para la evaluación subjetiva de la televisión de definición estándar (TVDS) e el entorno doméstico se indican en el Cuadro 3-2.

CUADRO 3-2

Condiciones de observación específicas para la evaluación subjetiva de sistemas digitales en el entorno doméstico

Condición	Descripción	Valores
a	Relación entre la distancia de observación y la altura de la imagen	6 H
b	Tamaño de la pantalla para una relación de formato de 4/3	De 25 a 29 pulgadas(1)
c	Tamaño de pantalla para una relación de formato de 16/9	De 32 a 36 pulgadas(1)
d	Norma de la pantalla	TVDS
e	Luminancia de cresta	200 cd/m^2
f	Iluminación ambiental sobre la pantalla (La luz ambiental que incide sobre la pantalla debe medirse perpendicularmente a ésta)	200 Lux

(1) Este tamaño de pantalla cumple las normas en materia de distancia de observación preferida (PVD, *preferred viewing distance*) para una PVD = 6 *H*.

A1-3 Métodos de evaluación

A1-3.1 Evaluación de la calidad de imagen básica

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

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

Se recomienda que en la evaluación se utilicen secuencias de al menos seis imágenes, más una adicional destinada a la capacitación antes del comienzo de la prueba. Las secuencias deben variar entre moderadamente críticas y críticas en el contexto de la aplicación de reducción de velocidad binaria que esté en consideración.

A lo largo de este Anexo, se hace énfasis en la importancia de comprobar los códecs digitales con secuencias de imágenes que sean críticas en el contexto de la reducción de la velocidad binaria en televisión. Parece por ello razonable preguntarse en qué medida es crítica una secuencia de imágenes particular para un objetivo determinado de reducción de la velocidad binaria, o si una secuencia es más crítica que otra. Una respuesta sencilla, aunque no especialmente útil, sería decir que «criticidad»

significa cosas muy distintas para diferentes códecs. Por ejemplo, podría ocurrir que una imagen fija que contuviera muchos detalles resultase crítica para un códec intratrama mientras que para un códec intercuadro, que es capaz de aprovechar similitudes de cuadro a cuadro, esa misma escena no representaría ninguna dificultad. Algunas secuencias que emplean textura móvil y movimiento complejo resultan críticas para toda clase de códecs, por lo que estos tipos de secuencias son los que más interesa generar o identificar. El movimiento complejo puede tomar la forma de movimientos que son predecibles para un observador pero no para los algoritmos de codificación, como por ejemplo un movimiento periódico tortuoso.

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

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

a) *Códecs intracampo de longitud de palabra fija*

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

b) *Códecs intercuadro de longitud de palabra fija*

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

c) *Códecs intracampo de longitud de palabra variable*

Se recomienda que estos códecs se prueben con material de secuencias de imágenes en movimiento, por las mismas razones que los códecs de longitud de palabra fija. Hay que tener en cuenta que debido a su codificación de longitud de palabra variable y su memoria intermedia asociada, estos códecs pueden distribuir dinámicamente la capacidad de bits de codificación a través de la imagen. Así por ejemplo, si en la mitad de una imagen se presenta un cielo sin rasgos especiales que no necesita muchos bits para su codificación, se ahorra capacidad para otras partes de la imagen que pueden así reproducirse con calidad elevada, incluso si son críticas. La conclusión importante de todo esto es que si una secuencia de imágenes resulta crítica para un códec de este tipo, habrá que detallar el contenido de cada

parte de la pantalla. Debe llenarse con textura en movimiento y estática, con tanta variación de color como se pueda y objetos con bordes nítidos de alto contraste. Al menos una secuencia del conjunto de prueba debe presentar ruido de origen justamente perceptible y por lo menos una debe ser sintética.

d) *Códecs intercuadro de longitud de palabra variable*

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

A1-3.2 Evaluación de la calidad de la imagen después de la posproducción

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

La metodología de la prueba que debe utilizarse es la variante II del método de doble estímulo con escala de calidad continua. Sin embargo, aquí la condición de referencia es la fuente sometida a la misma posproducción que las imágenes decodificadas. Si se considera ventajoso incluir una referencia de calidad inferior también ella deberá someterse a la misma posproducción.

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

Debido a las limitaciones de las posibilidades prácticas de tener que evaluar un códec con varias posproducciones, el número de secuencias de imágenes de prueba utilizadas puede ser como mínimo de tres, y una más disponible a efectos de demostración, por la imposición de tipo práctico de tener que evaluar probablemente un códec con varias posproducciones. La naturaleza de las secuencias dependerá de la tarea de posproducción que se estudie, pero debe variar entre moderadamente crítica y crítica en el contexto de reducción de la velocidad binaria de televisión y para el proceso que se considere. Para la evaluación de la cámara lenta puede servir una velocidad de visualización que sea la décima parte de la de origen.

A1-3.3 Evaluación de la característica de fallo

En la evaluación subjetiva de las degradaciones de las imágenes de códec debidas a imperfecciones del canal de transmisión o emisión, debe elegirse al menos cinco, pero preferiblemente más, proporciones de bits erróneos o condiciones de transmisión/emisión seleccionadas, con separación aproximadamente logarítmica y que abarquen la gama que provoca las degradaciones de códec desde «imperceptible» a «muy molesta».

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

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

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

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

A1-3.4 Características de fallo de la imagen según su contenido

En el Anexo 1 a la Parte 1 se describe el concepto general de características de fallo de la imagen según su contenido. Para aplicar este concepto al sistema de televisión de definición estándar, se debe utilizar el siguiente procedimiento.

A1-3.4.1 Definición de la criticidad

Conviene definir una medida determinada, llamada «criticidad», que representa las características del sistema de televisión digital en prueba y que se mide por un método objetivo. Como ejemplo de sistema de televisión digital se utiliza MPEG-2 MP@ML y la criticidad aplicada se basa en el método de entropía de cuantificador fijo descrito en la Recomendación UIT-T BT.1210.

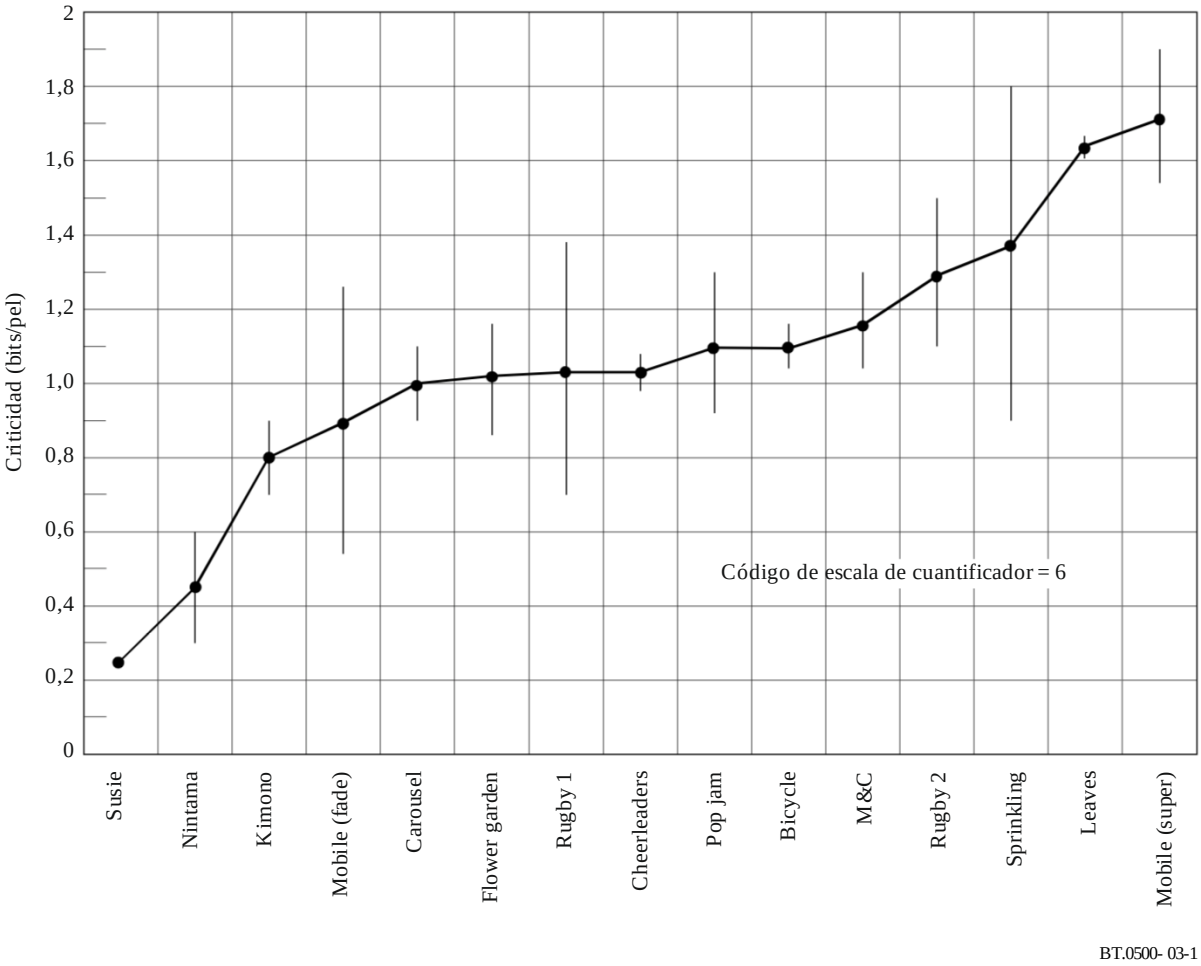
A1-3.4.2 Procedimiento para obtener las características de fallo de la imagen según su contenido

– *Fase 1*: Medición de la criticidad de las secuencias de prueba utilizadas en la evaluación subjetiva.

Se mide la criticidad de las secuencias de prueba utilizadas para la evaluación subjetiva descrita en la fase 3 siguiente. En la Fig. 3-1 se puede ver la desviación media y típica de cada secuencia en el sistema utilizado como ejemplo. La mayoría de las secuencias tienen una medida de criticidad de 0,8 a 1,4 bits/píxel. Algunas secuencias tienen una desviación típica grande porque el contenido de imagen varía considerablemente durante la secuencia.

FIGURA 3-1

Desviaciones media y típica de la criticidad de las secuencias de prueba

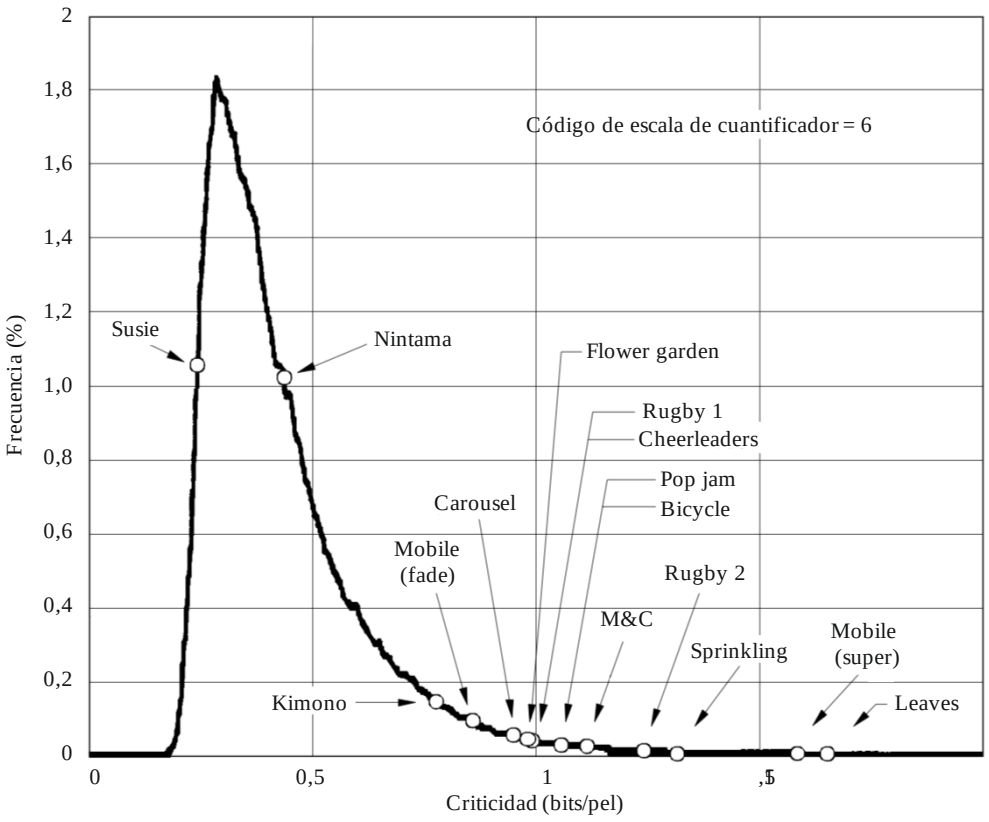


– *Fase 2*: Medición de la distribución de la criticidad de los programas de radiodifusión durante un largo periodo de tiempo.

La distribución de la criticidad de los programas de televisión transmitidos se mide durante un periodo de tiempo suficientemente largo, por ejemplo, una semana. En la Fig. 3-2 se presenta un ejemplo de la distribución medida durante una semana, 130 en total, de señales de radiodifusión NTSC, convertidas para la medición en las señales *Y*/*C* componentes. La frecuencia de aparición de criticidad en los programas de televisión se calculó cada 5 x 10^{-3} bits/píxel. En esta figura se presenta también la criticidad correspondiente a las secuencias se prueba utilizadas para la evaluación subjetiva.

FIGURA 3-2

Desviaciones de la criticidad de los programas transmitidos y criticidad de las secuencias de prueba



– *Fase 3*: Evaluación subjetiva de la calidad de imagen del sistema probado y obtención de la relación entre la criticidad y la calidad de imagen subjetiva.

La calidad de imagen del sistema de televisión digital se evalúa por el método de escala de calidad continua de doble estímulo (DSCQS, *double stimulus continous quality scale*). La relación entre la criticidad y las notas obtenidas en la evaluación se deriva combinando el resultado de la evaluación subjetiva y la criticidad obtenida en la fase 1. En la Fig. 3-3 se presenta la calidad de imagen del sistema utilizado como ejemplo a las velocidades binarias de 4, 6, 9 y 15 Mbit/s. La diferencia de calidad (% de DSCQS) en la Figura representa la degradación en relación con la secuencia de referencia original cuyos componentes tienen una relación 4:2:2. En la Fig. 3-4 se puede ver la relación entre la criticidad y la diferencia de calidad. En este ejemplo, se dio por supuesto una relación lineal entre criticidad y calidad de imagen y se derivaron las rectas de regresión utilizando el método de los mínimos cuadrados. En la Figura se ilustra la recta de regresión a cada velocidad binaria. En general, es posible aplicar una relación no lineal según los resultados de la evaluación.

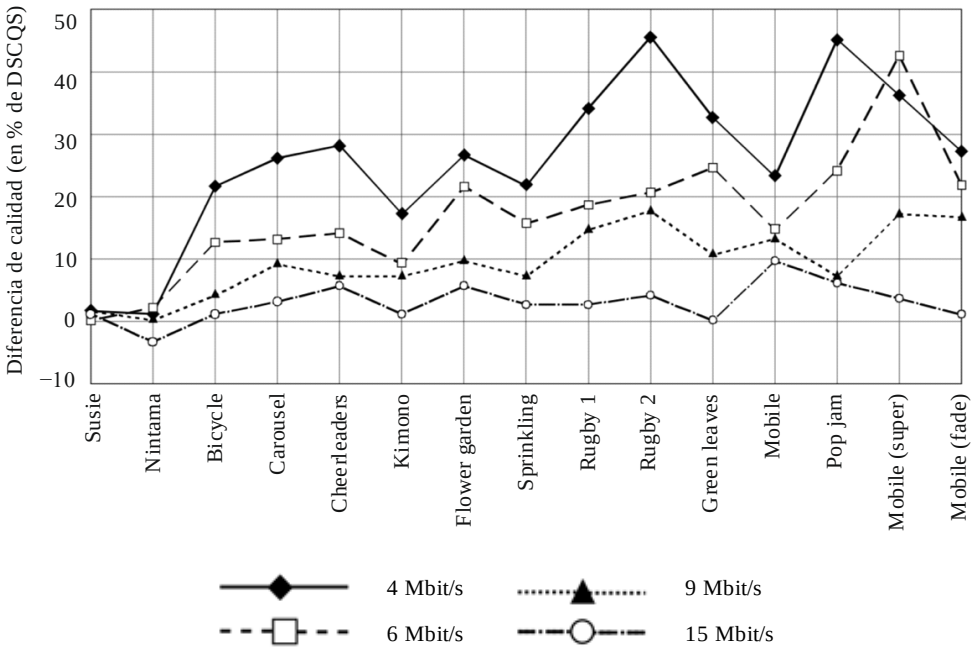
– *Fase 4*: Establecer las características de fallo de la imagen según su contenido (calidad en función de la frecuencia de aparición) combinando los resultados de la fase 3 (criticidad en función de la calidad) y de la fase 2 (criticidad en función de la frecuencia de aparición)

Combinando los resultados obtenidos en las fases 2 y 3 se obtienen las características de fallo de la imagen según su contenido, es decir, la distribución de la calidad de imagen de programas de televisión con codificación digital. La degradación de imagen en los programas de televisión

transmitidos se convierte en frecuencia acumulativa de aparición. En la Fig. 3-5 se presentan las características de fallo de la imagen según su contenido del sistema utilizado como ejemplo.

FIGURA 3-3

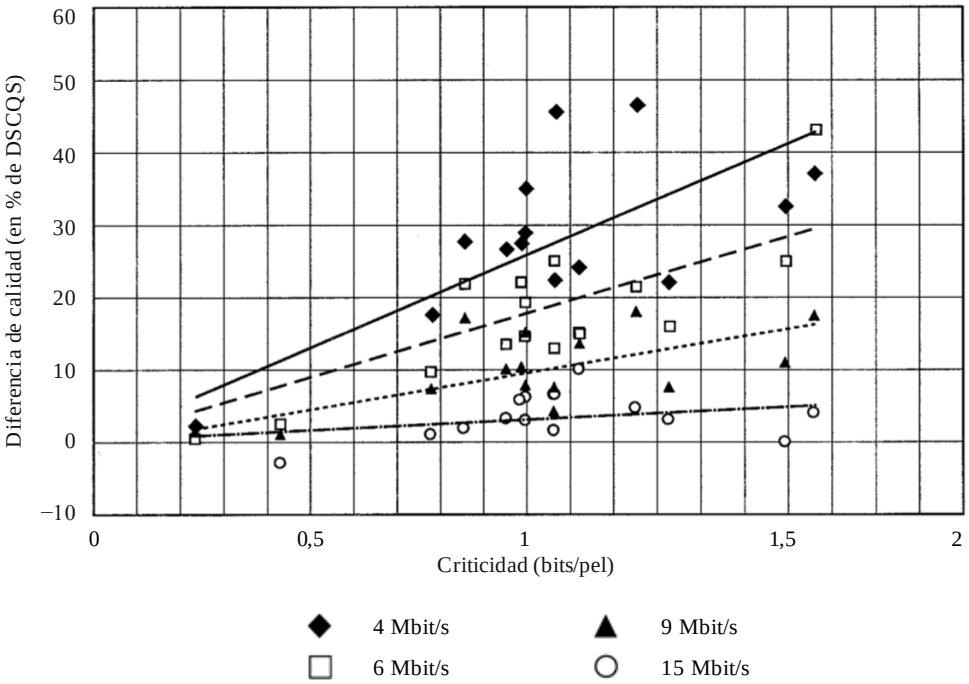
Resultado de la evaluación subjetiva (MP@ML en 6H)



BT.0500-03-3

FIGURA 3-4

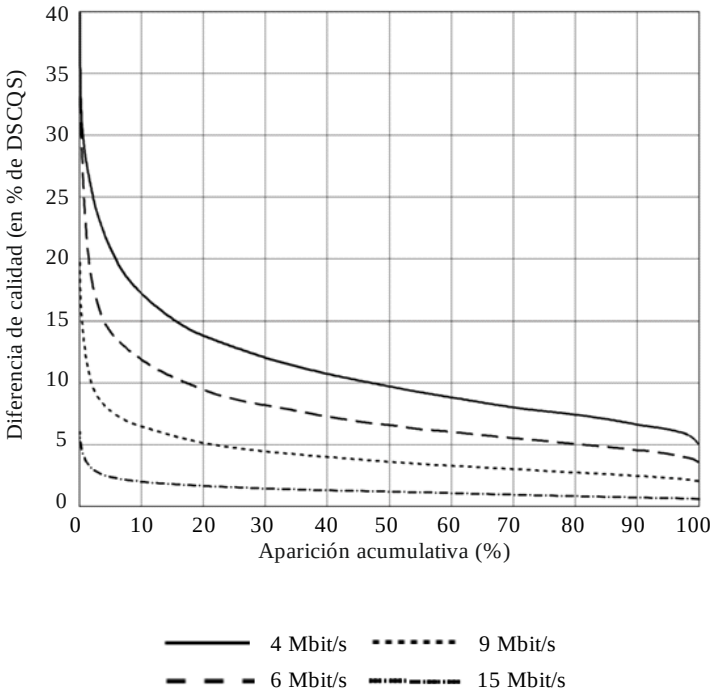
Relación entre la criticidad y la nota de evaluación (MP@ML en 6H)



BT.0500-03-4

FIGURA 3-5

Frecuencia acumulativa de aparición de degradación de imagen (MP@ML en 6H)



BT.0500- 03-5

A1-4 Observaciones relativas a la aplicación

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

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

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

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

Si el canal de radiodifusión se utiliza para entregar trenes de bits de programas múltiples o métodos de codificación jerárquica o ajustable por escalón, quizás sea necesario adaptar la metodología de evaluación para tener en cuenta lo siguiente:

– puede que el criterio de servicio aceptable no sea la transparencia de la codificación en la fuente; en vez, puede ser la capacidad del sistema de suministrar, con una velocidad binaria determinada, una alternativa viable para el servicio convencional. Por consiguiente, como sucede con la referencia en las pruebas de calidad, quizás sea conveniente utilizar el material que puede entregar un sistema convencional en condiciones de recepción típicas, en vez de entregarlo en forma digital sin compresión. Además, quizás sea conveniente utilizar material de prueba seleccionado para representar la gama de programas actuales y futuros (véase el Anexo 3 a la Parte 1). En las pruebas, las condiciones de observación serán las indicadas en la Parte 1 y en el § A1-2 de este Anexo, y el método general de prueba será el de doble estímulo con escala de calidad continua (véase el Anexo 2 a la Parte 2); y,

– es importante tener en cuenta la capacidad del sistema de mantener la integridad de cada tren de bits de programa en condiciones de carga total de canal y las degradaciones de la transmisión. Por consiguiente, en las pruebas de degradación, quizás resulte conveniente asegurar la carga total del canal y utilizar una gama de niveles de degradación seleccionada de manera que represente la gama de condiciones de recepción posibles (véase el Anexo 4 a la Parte 1). En las pruebas, las condiciones de observación serán las indicadas en la Parte 1 y en el § A1-2 de este Anexo, y el método de prueba general será el de doble estímulo con escala de degradación (véase el Anexo 1 a la Parte 2).

NOTA – Cuando se evalúan sistemas analógicos y digitales en el mismo contexto, es importante elegir un conjunto de materiales de prueba que refleje una dificultad equilibrada para los sistemas analógicos y digitales. En este caso y para profundizar el análisis, quizás resulte conveniente aplicar el método con escala multidimensional.

Anexo 2
a la Parte 3

Métodos de evaluación subjetiva de la calidad de las imágenes en sistemas de televisión de alta definición (TVAD)

A2-1 Condiciones de observación

A menos que se indique lo contrario en el Cuadro 3-3, las condiciones de observación deberían ser las descritas en el § 2 de la Parte 1.

CUADRO 3-3

Condiciones de observación para la evaluación subjetiva de la calidad de las imágenes TVAD

Condición	Descripción	Valores
a	Relación entre la distancia de observación y la altura de la imagen	3
b	Luminancia de cresta en la pantalla (cd/m^2)(1)	150-250
c	Relación entre el valor de luminancia de la pantalla inactiva y el valor de cresta(2)	≤ 0,02
d	Relación entre el valor de luminancia cuando se representa en la pantalla sólo el nivel de negro en una habitación completamente a oscuras y el valor correspondiente al blanco más intenso(3)	Aproximadamente 0,01
e	Relación entre el valor de luminancia del fondo situado detrás de la pantalla de imágenes y el valor de luminancia de cresta de la imagen	Aproximadamente 0,15
f	Iluminación procedente de otras fuentes(4)	Baja
g	Cromaticidad del fondo	D65
h	Ángulo subentendido por aquella parte del fondo que satisface las especificaciones anteriores(5). Este valor debería respetarse para todos los observadores	53° de altura × 83° de anchura
i	Colocación de los observadores	Dentro de ±30° horizontalmente desde el centro de la pantalla. El límite vertical está en estudio
j	Tamaño de la pantalla(6)	1,4 m (55 pulgadas)

(1) Valor de luminancia de cresta en la pantalla, correspondiente a la señal de vídeo con una amplitud del 100%.

(2) En este punto podría influir la iluminación de la habitación y la gama de contraste de la pantalla.

(3) El nivel de negro corresponde a la señal de vídeo con amplitud del 0%.

(4) La iluminación de la habitación será tal que satisfaga las condiciones c y e.

(5) Se recomienda un mínimo de 28° de altura × 48° de anchura.

(6) Deberán utilizarse valores mayores o iguales que ≥ 76,2 cm (30 pulgadas), si no se dispone de pantallas del tamaño especificado. Véase la Nota 3 de la Parte 1.

A2-2 Métodos de evaluación

Las evaluaciones subjetivas de la calidad global de una imagen de TVAD entregada por un sistema de emisión se debe efectuar utilizando un método de doble estímulo con escala de calidad continua (véase el Anexo 2 a la Parte 2) tomando como referencia la imagen de calidad de TVAD de estudio.

La evaluación del comportamiento respecto a fallos del sistema de emisión de TVAD se debe realizar utilizando un método de degradación con doble estímulo (véase el Anexo 1 a la Parte 2), tomando como referencia la imagen de TVAD de estudio o bien la de emisión sin degradación.

Al abordar la calidad de funcionamiento para todo tipo de contenido de programa y condiciones de transmisión que puedan presentarse en la práctica, se debe tener en cuenta la descripción de las características de fallo compuestas que aparecen en el Anexo 4 a la Parte 1.

Utilizando estos métodos, se debe tener cuidado en distinguir la influencia del formato de presentación cuando éste es distinto del formato del sistema básico (por ejemplo, cualquier conversión ascendente). Se considera que es adecuado y aplicable que las evaluaciones suplementarias se lleven a cabo utilizando diferentes presentaciones visuales con el objeto de tener en cuenta distintos formatos de presentación.

Algunos de los sistemas de emisión de TVAD pueden incluir la incorporación de un formato de televisión convencional («compatibilidad hacia atrás»). Así pues, es necesario evaluar, en términos de calidad de la imagen, la adecuación de las imágenes de televisión convencional incorporadas a las emisiones de TVAD. Por estos sistemas, se deben aplicar las condiciones de visualización y los métodos de evaluación que figuran en el Anexo 1 a la Parte 3.

Se deben aplicar los conceptos y procedimientos básicos descritos en el Anexo 1 a la Parte 3 para sistemas digitales de emisión de TVAD que emplean esquemas de reducción de velocidad binaria.

A2-3 Materiales de prueba

En el Informe UIT-R BT.2245 se enumera una serie relativamente grande de imágenes fijas y secuencias en movimiento. Estas imágenes se deben utilizar preferiblemente como material de prueba común para la evaluación de calidad de la TVAD.

Anexo 3
a la Parte 3

Evaluación subjetiva de la calidad de las imágenes alfanuméricas y gráficas en servicios de teletexto y otros servicios de texto similares

Introducción

Ciertos sistemas gestionan imágenes gráficas y alfanuméricas y las transmiten utilizando los códigos digitales adecuados. Las imágenes gráficas y alfanuméricas revisten un carácter particular, distinto del de las imágenes de televisión convencionales, y el proceso mental que interviene en su evaluación subjetiva puede ser diferente.

En la presente Recomendación se incluyen métodos para evaluar la calidad subjetiva de las imágenes de los actuales programas de televisión. Es preciso realizar estudios sobre la calidad de las imágenes alfanuméricas y gráficas que se utilizan en diversos servicios nuevos transmitidos por el canal de televisión; dichos servicios emplean códigos digitales para describir las imágenes alfanuméricas y gráficas. Algunos parámetros de transmisión tienen repercusiones sobre la calidad de las imágenes visualizadas: resolución de página (número de filas por página y número de caracteres por fila) en el caso de la codificación alfamosaica de teletexto, la resolución de células de carácter (número de

pixeles y líneas por células) en el caso de la codificación de juegos de caracteres dinámicamente redefinibles (JCDR (véase la Recomendación UIT-R BT.653)), y la resolución de imagen en el caso de la audiografía de difusión, del facsímil o del teletexto. Además, también deben tenerse en cuenta los errores de transmisión que pueden afectar a los códigos. Por lo tanto, es necesario efectuar mediciones de la calidad y determinar las relaciones entre la calidad objetiva y la subjetiva para estos parámetros.

Se han realizado estudios que muestran los diferentes aspectos necesarios para la evaluación de la calidad de esas imágenes, cuyas características son distintas de las imágenes de televisión convencional. Ciertos parámetros, como el formato de pixel, la resolución de célula de carácter, los espaciamientos, los colores y la disposición, tienen repercusiones sobre diversos atributos de calidad: legibilidad, calidad, comodidad, molestia, esfuerzo de lectura, fatiga y factores estéticos. Se consideran tres aspectos principales: las condiciones de observación, los métodos de evaluación y el contexto de evaluación.

Dada la importancia de sentar las bases de las evaluaciones subjetivas de la calidad de las imágenes alfanuméricas y gráficas, todos los informes de pruebas deberían contener descripciones íntegras de las configuraciones y los materiales de prueba, los observadores y los métodos utilizados.

A3-1 Condiciones de observación

En la Parte 1 se definen las condiciones de observación para las imágenes de televisión que corresponden a bajos niveles de iluminación en la sala. Es probable que las imágenes alfanuméricas y gráficas se observen también en las condiciones normales de iluminación. Así, se ha propuesto el estudio de un conjunto complementario de condiciones de observación: iluminación de 500 lux, luminancia máxima de la pantalla de 70 a 200 cd/m^2, relación de contraste de la pantalla de 30 a 50 y valor de 1/4 para la relación entre la luminancia de fondo (de las paredes de la sala) y la luminancia máxima de la pantalla. Debería considerarse también la distancia de observación (de 4 a 8 veces la altura de la imagen).

A3-2 Métodos de evaluación

Se han hecho numerosos estudios sobre los aspectos tipográficos. La mayoría de ellos han utilizado «medidas de calidad», tales como los umbrales de detección o reconocimiento, el índice de reconocimiento, la velocidad de lectura, etc. Muy pocos han utilizado las «medidas subjetivas», tradicionalmente usadas en la evaluación de las imágenes de televisión. Se considera que los nuevos sistemas transmitidos en los canales de televisión deben tener buena calidad (por ejemplo, porcentaje de reconocimiento de las letras superior al 95%). Se podrían utilizar eficazmente las escalas de calidad y degradación incluidas en la presente Recomendación, aunque es preciso estudiar hasta qué punto esas escalas se pueden referir a la legibilidad. Se ha tratado de hacer una comparación con los métodos de evaluación de la calidad para fonía (UIT-T) y se sugiere el estudio de una escala de «esfuerzo de lectura» de 5 notas.

En otro método se comparan los resultados de evaluaciones subjetivas realizadas utilizando dos escalas diferentes de 5 notas que se indican a continuación en el Cuadro 3-4.

CUADRO 3-4

Escalas de legibilidad y esfuerzo de lectura

Escala de calidad de legibilidad	Escala del esfuerzo de lectura
Legibilidad excelente	Ningún esfuerzo de lectura
Legibilidad buena	Atención necesaria, pero sin gran esfuerzo de lectura
Legibilidad aceptable	Esfuerzo de lectura moderado
Legibilidad mediocre	Esfuerzo grande de lectura
Legibilidad mala	Esfuerzo muy grande de lectura

Se consideró importante que el texto de cada nota fuese bien explícito. Los valores medios de las notas obtenidas con la escala de esfuerzo de lectura son en general más altos que los obtenidos con la escala de legibilidad y la gama de notas facilitada por los observadores es más alta en el caso de la escala del esfuerzo de lectura.

En otro experimento se utilizó la escala de calidad descrita en el § A3-4.1 de la Parte 2 para evaluar tanto la calidad como la legibilidad globales de un texto mecanografiado y transmitido por un sistema de televisión con un tipo de línea y una anchura de banda variables. En cada condición se utilizaron dos modelos, uno de los cuales era más complejo y exacto que el otro, aunque ambos se referían al concepto de adición de la «escala de degradación», y se observó que ambos reflejaban los efectos combinados de una definición horizontal y vertical limitada. También se efectuaron mediciones de la legibilidad sobre la base de la proporción de caracteres correctamente identificados. Sin embargo, sobre esa base el valor de la legibilidad siguió siendo elevado mientras que el de la calidad permaneció bajo, y resulta evidente que, en general, el primer criterio es menos útil.

En otro estudio se efectuaron comparaciones de calidad de funcionamiento y de métodos subjetivos de evaluación empleando textos impresos en papel con caracteres de anchura fija y de anchura variable. Los métodos subjetivos demostraron ser más sensibles. El mismo tipo de estudio se repitió con visualización en la pantalla de un tubo de rayos catódicos, utilizando esta vez únicamente métodos subjetivos. El empleo de estos métodos subjetivos ha permitido obtener resultados sobre los tamaños visualmente ópticos de las matrices fijas y variables.

A3-3 Contexto de la evaluación

En un nuevo método de evaluación del servicio se considera que es posible definir de manera precisa las actividades de los usuarios del servicio estudiado. Las estimaciones no se hacen según el método convencional de presentar imágenes y solicitar simplemente estimaciones subjetivas normalizadas. En lugar de ello, los observadores utilizan las imágenes presentadas como si estuvieran utilizando el servicio previsto y todas las evaluaciones se llevan a cabo en este contexto.

La simulación de la utilización del servicio no excluye el empleo de mediciones subjetivas convencionales; no obstante establece, para las evaluaciones subjetivas, un contexto más adecuado para el servicio previsto. Puede asimismo permitir la utilización de medidas objetivas del comportamiento del observador y el desarrollo de nuevas medidas subjetivas particularmente adecuadas al servicio y a los parámetros que se estudian. Por último, la simulación establece una base más segura para generalizar los resultados de las evaluaciones de laboratorio a las condiciones del servicio estudiado.

Anexo 4
a la Parte 3

Evaluación subjetiva de la calidad de las imágenes de los servicios multiprograma[5]

Introducción

Para la evaluación subjetiva de la calidad de uno de los programas comprimidos y codificados con velocidad binaria constante (CBR, *constant bit rate*) dentro de un servicio multiprograma, deberían utilizarse los procedimientos subjetivos detallados en los Anexos 1 ó 2 a la Parte 3 y el procedimiento descrito en el § A4-2 del presente Anexo.

Para la evaluación subjetiva de la calidad de cada uno de los programas comprimidos y codificados con velocidad binaria variable (VBR, *variable bit rate*), utilizando métodos tales como la multiplexación estadística o la codificación conjunta, dentro de un servicio multiprograma, deberían utilizarse los procedimientos subjetivos detallados en los Anexos 1 ó 2 de la Parte 3 y el procedimiento descrito en el § A4-3 del presente Anexo.

A4-1 Pormenores de las evaluaciones generales

– Las evaluaciones de la calidad de los canales de contenido temático deberían realizarse utilizando un material de prueba cuyo contenido y criticidad sean similares a los del material que normalmente se vaya a transmitir por esos canales.

– A fin de evaluar la calidad global percibida de programas cuya calidad «instantánea» varía durante un periodo de tiempo, deberían utilizarse los procedimientos descritos en los § A4-2 y A4-3.

– La valoración por escalas de los resultados de los sistemas en los que se utilizan referencias de baja calidad, de acuerdo con los comentarios que figuran en la descripción del método de escala de calidad continua de doble estímulo (DSCQS, *Double Stimulus Continuous Quality Scale Method*), deberían aplicarse y estudiarse con más detenimiento, para efectuar pruebas en las que se comparen servicios multiprograma con material de baja calidad.

A4-2 Procedimientos de evaluación subjetiva de las imágenes de los servicios multiprograma de velocidad binaria constante

La evaluación subjetiva de la calidad de las imágenes de cada programa de TVDS y TVAD se puede efectuar de manera independiente utilizando los métodos descritos en el Anexo 1 (TVDS) o el Anexo 2 (TVAD) a la Parte 3. Para la evaluación de la calidad básica del sistema, debería utilizarse el método de prueba general DSCQS (descrito en el Anexo 2 a la Parte 2). Para la evaluación de programas con degradaciones de transmisión, debería utilizarse el método de escala de degradación con doble estímulo (DSIS, *Double Stimulus Impairment Scale*) de prueba general (descrito en el Anexo 1 a la Parte 2).

---

[5] Incluido el término «multiplexación estadística» o los servicios de «multiplexación estadística».

A4-3 Procedimientos de evaluación subjetiva de las imágenes de los servicios multiprograma de velocidad binaria variable

La evaluación subjetiva de la calidad de imagen de los programas de TVDS y TVAD codificados con velocidad binaria variable puede efectuarse de manera independiente utilizando el método DSCQS. También reviste especial importancia la selección de los materiales de prueba, ya que la calidad de la imagen puede depender del contenido de imagen de todos los programas multiplexados.

Anexo 5
a la Parte 3

Observación especializada de la calidad de las imágenes de los sistemas destinados a la proyección digital de imágenes digitales en pantalla grande[6] en cines

A5-1 Introducción

En los últimos años, la observación especializada se ha empleado frecuentemente para efectuar rápidas comprobaciones de la calidad de funcionamiento de un proceso de vídeo genérico.

El presente Anexo describe un método de prueba por observación especializada que permitirá una coherencia de resultados obtenidos en diferentes laboratorios al recurrir a los servicios de un número limitado de observadores especializados.

A5-2 Motivos del nuevo método basado en la «observación especializada»

Conviene poner de relieve las ventajas que resultan de la aplicación de la metodología propuesta.

En primer lugar, una prueba de evaluación subjetiva formal suele requerir el empleo de por lo menos 15 observadores elegidos entre los «no especializados», la realización de pruebas prolongadas y la búsqueda continua de nuevos observadores. Tal número de observadores es necesario para lograr la sensibilidad necesaria de modo que los sistemas sometidos a prueba puedan diferenciarse y clasificarse, o juzgarse equivalentes de manera fiable.

En segundo lugar, al recurrir a observadores no especializados, las formas tradicionales de prueba pueden no revelar diferencias que podrían resaltar en una exposición más prolongada, incluso para ojos no expertos.

En tercer lugar, las evaluaciones provisionales generalmente establecen medidas de calidad (o diferencias en la calidad), pero no identifican directamente los objetos u otras manifestaciones materiales a las que dichas medidas se refieren.

La metodología aquí propuesta trata de dar solución a estos tres problemas.

---

[6] Las imágenes digitales en pantalla grande (LSDI, *large screen digital imagery*) integran una familia de sistemas de imágenes digitales para programas de la índole de representaciones dramáticas, obras de teatro, acontecimientos deportivos, conciertos, eventos culturales, etc., incluidas desde la captura hasta la representación en pantalla grande con calidad de alta resolución en salas de cine, salones y otros lugares debidamente equipados.

A5-3 Definición de especialistas

A los efectos del presente Anexo, un «observador especializado» es una persona que conoce el material empleado para la evaluación, sabe «qué mirar», eventualmente puede estar bien informado sobre los detalles del algoritmo utilizado para procesar el material vídeo que ha de evaluarse. En todos los casos, el «observador especializado» es una persona con larga experiencia en el terreno de la investigación de la calidad, alguien profesionalmente vinculado al tema específico de la prueba. Por ejemplo, cuando se organiza una serie de pruebas de «observación especializada» de un material dado de LSDI, deberían elegirse expertos en la producción o postproducción de películas o en la producción de contenido vídeo de alta calidad (por ejemplo, camarógrafos, retocadores de color, etc.); la selección ha de efectuarse teniendo en cuenta la posibilidad de establecer juicios subjetivos únicos sobre la calidad de imagen de LSDI y los artilugios de compresión.

A5-4 Selección de evaluadores

Toda prueba de observación especializada es una serie de evaluaciones basadas en las opiniones de evaluadores y en la que se emiten juicios sobre la calidad visual y/o la degradación aparente.

El grupo básico de expertos está formado por cinco a seis personas. Este número reducido facilita la tarea de elegir los evaluadores y permite llegar a decisiones más rápidas.

En función de las necesidades del experimento, se acepta la utilización de más de un grupo básico de expertos, reunidos en un conjunto combinado más amplio (por ejemplo, provenientes de diferentes laboratorios).

Se entiende que los especialistas tienen tendencia a mejorar la clasificación cuando están probando su propia tecnología, por lo que debería evitarse la inclusión de personas directamente involucradas en el desarrollo del sistema sometido a prueba.

Todos los evaluadores deberían pasar por un examen de agudeza visual, normal o corregida (prueba de Snellen), así como de visión cromática normal (prueba de Ishihara).

A5-5 Material de prueba

Los materiales por probar deben seleccionarse de tal manera que sirvan de muestra de toda la gama de valores de producción y niveles de dificultad previstos en el contexto real en el cual los sistemas bajo prueba habrán de utilizarse. La selección debería favorecer los materiales más exigentes, aunque sin excesos. Por lo general, deberían emplearse de 5 a 7 secuencias de prueba.

El método de selección de materiales puede variar también en función de la aplicación para la cual el sistema bajo prueba se ha diseñado.

En tal sentido, no se dan más indicaciones aquí sobre las reglas de selección del material de prueba, dejando la decisión al autor de la concepción de la prueba, sobre la base de las consideraciones antes mencionadas.

A5-6 Condiciones de observación

Las condiciones de observación, que deberán describirse exhaustivamente en el informe sobre la prueba, han de ajustarse al Cuadro 3-5 y mantenerse de manera constante durante la prueba.

CUADRO 3-5

Panorama de las condiciones de observación

Condiciones de observación	Valores	
	Mínimo	**Máximo**
Dimensión de la pantalla (m)	6	16
Distancia de observación[(1)]	1,5 H	2 H
Luminancia del proyector (pantalla central, blanco máximo)	34 cd/m^2	48 cd/m^2
Luminancia de la pantalla (fuera del proyector)		<1/1 000 de la luminancia del proyector

(1) Ha de emplearse la presentación en «mariposa» cuando la distancia de observación sea menor que 1,5 H. En caso de emplearse una presentación «por yuxtaposición», la distancia de observación debería acercarse más al valor de 2 H.

A5-7 Metodología

A5-7.1 Series de evaluación

Cada serie de evaluación (definida como el conjunto de sesiones de prueba de un grupo determinado de observadores) debería comprender dos fases (por ejemplo, Fase I y Fase II).

A5-7.1.1 Fase I

La Fase I consiste en una prueba subjetiva formal realizada en un entorno controlado (véase el § A5-6), gracias a la cual se obtendrán resultados válidos, sensibles y repetibles. En esta fase, los especialistas clasifican cada uno el material presentado, empleando la escala de clasificación que se describe más adelante. Los miembros del equipo no deben debatir entre sí lo que estén viendo ni controlar las presentaciones. En el curso de esta fase, los especialistas no deberían conocer el sistema de codificación sometido a prueba ni el orden de presentación del material que se prueba. El material bajo prueba debería presentarse de manera aleatoria, para evitar toda parcialidad en el juicio.

A5-7.1.1.1 Presentación del material

El método de presentación combina elementos de doble estímulo simultáneo para evaluación continua (SDSCE) (véase el Anexo 6 a la Parte 2) y el método de escala de calidad continua de doble estímulo (DSCQS) (véase el Anexo 2 a la Parte 2). Es posible referirse a él como método de doble estímulo simultáneo.

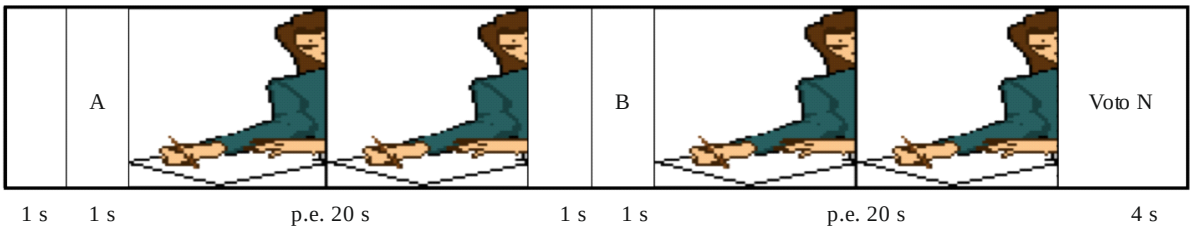
Con arreglo al método SDSCE, en cada prueba se presentará en una pantalla dividida el material correspondiente a dos imágenes. En la mayoría de los casos, una de las imágenes de origen será la referencia (es decir, la imagen fuente) y la otra la imagen de prueba; en otros casos, las dos imágenes se habrán extraído de la imagen de referencia. La referencia estará constituida por el material de origen presentado de manera transparente (es decir, sin someterlo a otra compresión que la correspondiente al medio de grabación de la fuente). El material de prueba será el material de fuente procesado a través de uno de los sistemas bajo prueba. La velocidad binaria y/o el nivel de calidad corresponderán a las especificaciones del guion de la prueba. A diferencia del método SDSCE, los observadores no conocerán las condiciones representadas por ambos miembros del par de imágenes.

La presentación en pantalla dividida podrá efectuarse mediante el método tradicional de división de la pantalla sin efecto de simetría, o bien mediante la técnica «en mariposa», en la que la imagen de la derecha de la pantalla es su equivalente especular. Puesto que se presentarán imágenes completas a lo ancho, sólo la mitad de cada una podrá observarse a la vez. En cada presentación, en cada lado de la pantalla se presentará la misma mitad de la imagen.

Con arreglo al método DSCQS, el par de imágenes se presenta dos veces sucesivamente, una vez para familiarizarse con la imagen y evaluarla, y otra vez para confirmar la impresión y clasificar. Cada secuencia durará entre 15 y 30 s. Es posible rotular cada secuencia al comienzo de cada fragmento, para asistir a los evaluadores (véase la Fig. 3-6 como ejemplo de división de pantalla sin efecto especular).

FIGURA 3-6

Ejemplo de división de pantalla sin efecto especular

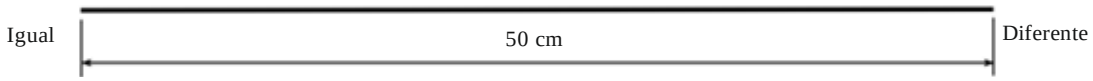


BT.0500- 03-6

A5-7.1.1.2 Escala de evaluación

El criterio de aceptación en las aplicaciones de LSDI es que la imagen de prueba (es decir, comprimida) no pueda distinguirse de la referencia. Para evaluar los sistemas sometidos a prueba pueden emplearse varios métodos corrientes de clasificación. Uno de los métodos que se propone es el basado en las escalas de comparación de estímulo (véase el Anexo 4 a la Parte 2). Un ejemplo concreto de escala es la de IGUAL-DIFERENTE, de tipo no categórico (continuo) que se describe en el § A4-4.2 del Anexo 4 a la Parte 2:

FIGURA 3-7



BT.0500- 03-7

A5-7.1.1.3 Sesiones de evaluación

El proceso de evaluación puede implicar más de una sesión, en función del número de condiciones de prueba, y deberá comprender dos tipos de pruebas: pruebas iniciales y pruebas de confirmación. En una prueba inicial, una mitad de la pantalla muestra la referencia, mientras que la otra mitad muestra la prueba. En una prueba de confirmación ambas mitades muestran la referencia. La prueba de confirmación tiene por objeto medir la posibilidad de juicio parcial.

Para cada sistema probado, se requieren en cada secuencia de prueba las siguientes pruebas iniciales:

CUADRO 3-6

Flanco izquierdo de la pantalla	Flanco derecho de la pantalla
Referencia de la mitad de la izquierda	Prueba de la mitad de la izquierda
Referencia de la mitad de la derecha	Prueba de la mitad de la derecha
Prueba de la mitad de la izquierda	Referencia de la mitad de la izquierda
Prueba de la mitad de la derecha	Referencia de la mitad de la derecha

De preferencia, cada uno de los casos descritos debería repetirse por lo menos dos veces. Para cada sistema se requiere, en cada secuencia de prueba, las siguientes pruebas de confirmación:

CUADRO 3-7

Flanco izquierdo	Flanco derecho
Referencia de la mitad de la izquierda	Referencia de la mitad de la izquierda
Referencia de la mitad de la derecha	Referencia de la mitad de la derecha

También aquí es preferible efectuar, para cada uno de los casos descritos, por lo menos dos repeticiones.

Las pruebas deberían consistir en sesiones de más de una hora de duración, con pausas de descanso de 15 min. Las pruebas iniciales y de confirmación resultantes de la combinación de secuencias de códec y de prueba deberían dividirse en sesiones según un criterio de asignación seudoaleatorio. Aunque sea más complicado, vale la pena imponer algunas restricciones a este proceso. Por ejemplo, en caso de efectuarse cuatro sesiones, podría asignarse aleatoriamente cada una de las pruebas iniciales a un códec dado y las secuencias de pruebas a una posición determinada aleatoriamente en una de estas sesiones. Este método tiene la ventaja de asegurar una distribución de las pruebas iniciales del sistema a lo largo de toda la serie de pruebas.

A5-7.1.1.4 Procesamiento de los resultados de las pruebas

Para una prueba inicial dada, el resultado de la prueba es la distancia entre el extremo «IGUAL» de la escala y la marca establecida por el observador, expresada en valores de 0 a 100 de la escala. Los resultados se analizarán como nota media de opinión (MOS), método que se empleará para establecer una clasificación de los sistemas probados. Según el número de observaciones efectuadas por sistema (observadores × secuencias de prueba × repeticiones), los datos pueden someterse a un análisis de varianza (ANOVA)[7]. El resultado de pruebas de confirmación puede utilizarse para derivar una diferencia de evaluación básica de «azar».

[7] Un total de 10-20 observaciones en la condición de orden inferior de interés es suficiente para aplicar tratamientos estadísticos inferenciales del tipo ANOVA.

A5-7.1.2 Fase II

Uno de los principales objetivos de la Fase II es ajustar el orden relativo de los resultados de la Fase I, cuya precisión y fiabilidad puede reducirse debido al número limitado de observadores y/o de evaluaciones efectuadas. Otro objetivo importante es hacer variar las observaciones en cuanto a las características de percepción de las imágenes en las que se hubieran basado las evaluaciones de la Fase I.

En esta parte, el grupo de expertos analiza el material presentado. Aquí están autorizados a comentar el material presentado, repetir su presentación, total o parcialmente, todas las veces que sea necesario para su análisis y/o demostración, y a elaborar por consenso una evaluación y descripción de lo que hayan visto. Si los observadores especializados lo desean, puede hacerse una «presentación trucada», empleando por ejemplo modalidades tales como movimiento lento, imágenes consecutivas y cuadros fijos. Estas técnicas requerirán cierta interacción y la intervención del director de la prueba.

A5-7.1.2.1 Agrupamiento del material sometido a prueba

Para llevar a cabo la prueba de la Fase II adecuadamente, es necesario agrupar el material sometido a prueba según su contenido, en lo que se denomina conjunto básico de observación para expertos (BES), o sea, todas las secuencias codificadas obtenidas de la misma secuencia fuente deben agruparse y ordenarse, de conformidad con los resultados de la Fase I.

Los materiales de prueba se ordenarán desde los valores inferiores de MOS hasta los valores más altos. El número de BES corresponderá al número de secuencias utilizadas en la prueba.

A5-7.1.2.2 Subsesión de prueba de observación especializada básica

La subsesión de prueba de expertos de observación básica (BEV) es una sesión de debate, en la que los expertos examinan todo el material incluido en un BES; una tarea consiste en confirmar o modificar el orden de clasificación resultante de la prueba formal de la Fase I. Por consiguiente, ha de confirmarse o modificarse la relativa visibilidad de las diferencias.

A5-7.1.2.3 Plan de la Fase II

En la Fase II, deben llevarse a cabo todas las BEV. Los expertos sabrán que el orden de presentación es el resultado de la clasificación de la Fase I. Los expertos no tendrán conocimiento de las relaciones entre los sistemas propuestos y su clasificación.

La Fase II se llevará a cabo como un esfuerzo colectivo que dé por resultado opiniones de consenso entre los evaluadores.

Antes de iniciarse la Fase II, se encargará a los evaluadores, eventualmente mediante un texto escrito, a ejecutar las tareas siguientes:

– Observar el material en cada BEV.
– Debatir la clasificación del material en cada BEV; en caso de desacuerdo del grupo, definir un nuevo orden de clasificación.
– Comentar cada caso, con inclusión de observaciones de detalle sobre la característica de las diferencias observadas, en caso de existir.
– Documentar la clasificación, así como sus comentarios y observaciones.

El director de la prueba tendrá a su cargo la responsabilidad de reunir todos los comentarios de los grupos y señalar las discrepancias. Mientras prosiguen las pruebas, los resultados de las Fases I y II de los distintos grupos se mantendrán en secreto, para no influir a los grupos siguientes. En la medida de lo posible, el director de la prueba está autorizado a identificar las divergencias y apoyar la solución de las mismas mediante pruebas adicionales para los resultados controvertidos. El objetivo de esta última medida es lograr un consenso general.

A5-8 Informe

El informe final de la prueba estará a cargo del director de la misma.

Dicho Informe comprenderá la siguiente información:

– resultados de la Fase I (con inclusión de cuadros de MOS, así como los resultados de los análisis estadísticos que corresponda);
– comentarios de los expertos durante la Fase II;
– comentarios sobre toda reclasificación;
– cualquier información pertinente sobre condiciones de observación, características de la señal de entrada, procesamiento de la señal, características del proyector, ajuste del proyector, cromaticidad, selección de observadores y condiciones de la prueba;
– una caracterización completa de la calidad de funcionamiento del dispositivo de visualización (tiempo medio entre fallos, etc.);
– resumen y conclusiones.

Anexo 6
a la Parte 3

Metodología para la evaluación subjetiva de la calidad de vídeo en aplicaciones multimedios

A6-1 Introducción

Un gran número de países ha empezado a instalar sistemas de radiodifusión digital que permitirán la distribución de aplicaciones de radiodifusión de datos y multimedios que comprenden vídeo, audio, imagen fija, texto y gráficos.

Se necesitan métodos de evaluación subjetiva normalizados con objeto de especificar los requisitos de calidad de funcionamiento y de verificar si las soluciones técnicas consideradas para cada aplicación son las adecuadas. Las metodologías subjetivas son necesarias porque proporcionan mediciones que permiten a la industria anticiparse de manera más directa a las reacciones de los usuarios finales.

El sistema de radiodifusión necesario para distribuir aplicaciones multimedios es muy diferente del que se usa actualmente: se accede a la información a través de receptores fijos y/o móviles, la velocidad de cuadro puede ser fija o variable; hay una amplia gama de posibles tamaños de imagen (SQCIF a TVAD); el vídeo se asocia típicamente al audio, texto y/o sonido intercalados; el vídeo se puede procesar con códecs de vídeo avanzado; y la distancia de observación preferida depende en gran medida de la aplicación.

Los métodos de evaluación subjetiva especificados en la Parte 2 deben aplicarse en este nuevo contexto. Asimismo, pueden llevarse a cabo investigaciones de sistemas multimedios mediante nuevas metodologías para ajustarse a los requisitos de usuario de las características del dominio multimedios.

En este Anexo se describe el proceso de evaluación subjetiva no interactiva de la calidad de vídeo para aplicaciones multimedios. Dichos métodos pueden aplicarse para diferentes propósitos, que incluyen, sin ser exhaustivos: selección de algoritmos, clasificación de la calidad de funcionamiento de sistemas audiovisuales y evaluación del nivel de calidad de vídeo durante una conexión audiovisual.

A6-2 Características comunes

A6-2.1 Condiciones de observación

En el Cuadro 3-8 se enumeran las condiciones de observación recomendadas. El tamaño y el tipo de visualización utilizados deben ser adecuados para la aplicación sobre la que se está investigando. Como deben emplearse varias tecnologías de visualización en las aplicaciones multimedios, es necesario comunicar toda información relevante sobre la visualización usada en la evaluación (por ejemplo, fabricante, modelo y especificaciones).

Cuando se haga uso de sistemas basados en PC para presentar las secuencias, deben señalarse asimismo las características de los sistemas (por ejemplo, la tarjeta de visualización de vídeo).

En el Cuadro 3-9 se muestra un ejemplo del registro de datos para la configuración del sistema multimedios sometido a prueba.

Si las imágenes de prueba se obtienen por medio de una combinación decodificador-reproductor específica, hay que separar las imágenes de la interfaz gráfica patentada para lograr una visualización anónima. Ello es necesario para asegurar que la evaluación de la calidad no se ve influenciada por el conocimiento del entorno original.

Cuando los sistemas evaluados en una prueba utilizan un formato de imagen reducido, como CIF, SIF o QCIF, las secuencias deben mostrarse en una ventana de la pantalla de visualización. El color del fondo de la pantalla debe ser gris al 50%.

CUADRO 3-8

Condiciones de observación recomendadas empleadas en evaluaciones de calidad de sistemas multimedios

Parámetro	Ajuste
Distancia de observación(1)	Limitado: 1-8 H Ilimitado: basado en las preferencias del espectador
Valor de la cresta de luminancia en la pantalla	70-250 cd/m^2
Relación entre la luminancia de la pantalla inactiva y la luminancia de cresta	≤ 0,05
Relación entre la luminancia de la pantalla, cuando se presenta únicamente el nivel de negro en una habitación totalmente oscura, y la correspondiente al blanco de cresta	≤ 0,1
Relación entre la luminancia de fondo detrás de la pantalla de imágenes y la luminancia de cresta de la imagen(2)	≤ 0,2
Cromaticidad del fondo(3)	D_{65}
Iluminación de fondo de la habitación(2)	≤ 20 lux

(1) La distancia de observación depende en general de la aplicación.

(2) Este valor indica un ajuste que permite la máxima detectabilidad de las distorsiones; para algunas aplicaciones se permiten valores superiores o vienen determinados por la aplicación.

(3) Para pantallas de PC, la cromaticidad del fondo debe aproximarse lo máximo posible a la de «punto blanco» de visualización.

CUADRO 3-9

Configuración del sistema multimedios sometido a prueba

Parámetro	Especificación
Tipo de visualización	
Tamaño de visualización	
Tarjeta de visualización de vídeo	
Fabricante	
Modelo	
Información de la imagen	

A6-2.2 Señales fuente

La señal fuente proporciona la imagen de referencia directamente, así como la entrada para el sistema que se está probando. La calidad de las secuencias fuente debe ser la mejor posible. Como referencia, la señal de vídeo debe grabarse en archivos multimedios usando yuv (formatos 4:2:2, 4:4:4) o RGB (24 ó 32 bits). Cuando el experimentador está interesado en comparar los resultados de diferentes laboratorios, es preciso utilizar un conjunto de secuencias común para eliminar fuentes adicionales de variación.

A6-2.3 Selección de materiales de prueba

El número y tipo de escenas de prueba resultan críticos a la hora de interpretar los resultados de la evaluación subjetiva. Algunos procesos pueden provocar el mismo nivel de degradación para la mayor parte de las secuencias. En tales casos, los resultados obtenidos con un número pequeño de secuencias (por ejemplo, dos) deben proporcionar una evaluación representativa. No obstante, los nuevos sistemas a menudo tienen una repercusión que depende en gran medida de la escena o del contenido de la secuencia. En tales casos, el número y tipo de escenas de prueba debe escogerse de modo que pueda generalizarse de manera razonable la programación normal. Asimismo, el material debe seleccionarse de modo que sea «crítico, pero dentro de unos límites razonables» para el sistema que se está probando. «Dentro de unos límites razonables» implica que la escena podría todavía formar parte del contenido normal de la programación de televisión. Las características de percepción espacial y temporal de una escena pueden proporcionar una indicación útil de la complejidad de la misma. En el Anexo 6 a la Parte 1 se presentan con más detalle las mediciones de características de percepción espacial y temporal.

A6-2.4 Gama de condiciones y anclaje

Dado que la mayoría de los métodos de evaluación son sensibles a variaciones en la gama y distribución de las condiciones vistas, las sesiones de decisión deben incluir las gamas completas de los factores variados. Sin embargo, esto podría aproximarse mediante una gama más restringida presentando también algunas condiciones que corresponderían a los extremos de las escalas. Pueden representarse como ejemplos e identificarse como extremos (anclaje directo), o distribuirse a través de la sesión y no identificarse como extremos absolutos (anclaje indirecto). En la medida de lo posible, debe emplearse una gran gama de valores de calidad.

A6-2.5 Observadores

El número de observadores tras el análisis debe ser de al menos 15; no expertos, en el sentido de que en su trabajo normal no están directamente interesados en la calidad de imagen y de que no son evaluadores con experiencia. Antes de una sesión, se debe analizar a los observadores para lograr (o corregir) la agudeza visual normal en el gráfico de Snellen o Landolt, y la visión en color normal, usando gráficos especialmente seleccionados (por ejemplo, Ishihara).

El número de evaluadores necesario depende de la sensibilidad y la fiabilidad del procedimiento de prueba adoptado, así como del tamaño anticipado del efecto buscado.

En los experimentos se debe incluir el mayor número posible de detalles acerca de las características de sus paneles de evaluación para facilitar una posterior investigación de este factor. Entre los datos que podrían proporcionarse figuran los de categoría laboral (por ejemplo, empleado de organización de radiodifusión, estudiante de universidad, empleado de oficina), sexo y edad.

A6-2.6 Diseño experimental

Se deja a criterio del experimentador la selección de un diseño experimental con el que satisfacer los objetivos de coste y precisión específicos. Se recomienda incluir al menos dos reiteraciones (es decir, repeticiones de condiciones idénticas) en el experimento. Las reiteraciones permiten calcular la fiabilidad individual de cada sujeto y, si fuera necesario, descartar resultados no fiables de algunos sujetos. Además, las reiteraciones aseguran que los efectos del aprendizaje en una prueba quedan compensados en cierta medida. Se obtiene una mejora adicional en el tratamiento de los efectos del aprendizaje mediante unas pocas «presentaciones ficticias» al comienzo de cada sesión. Estas condiciones deben elegirse de modo que sean representativas de las presentaciones que se van a mostrar más tarde durante la sesión. Las presentaciones preliminares no se han de tener en cuenta en el análisis estadístico de los resultados de la prueba.

La duración de una sesión, es decir, de una serie de presentaciones, no debe ser más de media hora.

Cuando se prueban varias escenas o algoritmos, el orden de presentación de las escenas o algoritmos debe aleatorizarse. El orden aleatorio puede modificarse para garantizar que las mismas escenas o los mismos algoritmos no estén próximos temporalmente (es decir, consecutivamente).

A6-3 Métodos de evaluación

Se puede examinar la calidad de funcionamiento del vídeo de los sistemas multimedios mediante las metodologías descritas en la Parte 2.

El método de evaluación subjetiva de calidad de vídeo multimedios (SAMVIQ) aprovecha las características del dominio multimedio y puede usarse para evaluar la calidad de funcionamiento de sistemas multimedios.

Anexo 7
a la Parte 3

Métodos de evaluación subjetiva de los sistemas de televisión 3D estereoscópica

A7-1 Dimensiones de la evaluación (perceptual)

La televisión en 3D estereoscópica explota las características del sistema de visión binocular humana, recreando las condiciones que producen la percepción de la profundidad relativa de los objetos en el campo visual. El principal requisito de la tecnología actualmente disponible para la composición de una imagen estereoscópica es la captura de al menos dos vistas de la escena tomadas por dos cámaras alineadas horizontalmente. Las imágenes de los objetos de la escena tendrán posiciones relativas diferentes en función de la visión de que se trate, a saber, visión izquierda y visión derecha. La diferencia entre las posiciones relativas de las dos visiones es lo que comúnmente se denomina disparidad de imágenes (o paralaje) y normalmente se expresa en píxeles, distancia física (retiniana) o con una medición relativa (por ejemplo, porcentaje de la anchura de la pantalla). La disparidad de imágenes debe distinguirse de la disparidad angular (retiniana). De hecho, la misma información de disparidad de imagen produce distintas disparidades angulares (retinianas) con distintas distancias de visualización. La magnitud y dirección de la percepción de profundidad se basa en la magnitud y dirección de las disparidades retinianas resultado de la imagen estereoscópica.

Los factores de evaluación generalmente aplicados a las imágenes de televisión monoscópica, tales como resolución, reproducción del color, representación del movimiento, calidad global, nitidez de la imagen, etc. también pueden aplicarse a los sistemas de televisión estereoscópica. Además, existen numerosos factores específicos de los sistemas de televisión estereoscópica. Entre ellos se incluyen factores como la resolución en profundidad, que es la resolución espacial en el sentido de la profundidad, el movimiento en profundidad, que determina si el movimiento en la dirección de la profundidad se reproduce sin discontinuidades ni distorsiones espaciales. Dos ejemplos bien conocidos de esto último son el *efecto teatro de marionetas,* que se produce cuando los objetos se perciben anormalmente grandes o pequeños, y el *efecto papel de cartón* (o aplanamiento de la profundidad de las imágenes), cuando los objetos se perciben de forma estereoscópica pero con una menor profundidad que hace que parezcan anormalmente delgados.

Pueden identificarse tres dimensiones básicas de la percepción, que en su conjunto afectan a la calidad de experiencia de un sistema estereoscópico: *calidad de la imagen, calidad de profundidad y confort visual*. Algunos investigadores consideran que el impacto psicológico de las tecnologías de representación estereoscópica también podría medirse mediante conceptos más generales, como la *naturalidad* o el *sentido de presencia*.

A7-1.1 Dimensiones perceptuales primarias

Calidad de la imagen, que hace referencia a la calidad de imagen que proporciona el sistema. Es el principal parámetro que determina la calidad de funcionamiento de un sistema de vídeo. La calidad de la imagen está principalmente afectada por parámetros técnicos y otros errores introducidos, por ejemplo, por procesos de codificación y/o transmisión

Calidad de la profundidad, que hace referencia a la capacidad del sistema para ofrecer una sensación mejorada de profundidad. La presencia de factores monoculares, como perspectiva lineal, imagen borrosa, gradientes, etc. ofrece sensaciones de profundidad incluso en imágenes 2D. No obstante, las imágenes 3D estereoscópicas contienen además información de disparidad que proporciona información de profundidad que mejora la sensación de profundidad en comparación con los sistemas 2D.

Confort visual, que se refiere a la sensación subjetiva de comodidad/incomodidad que puede asociarse a la visión de imágenes estereoscópicas. Las imágenes estereoscópicas que hayan sido capturadas de forma incorrecta o sea representadas de forma inadecuada pueden generar una notable incomodidad.

A7-1.2 Dimensiones adicionales de la percepción

Naturalidad, que se refiere a la percepción de la imagen estereoscópica como una representación fiel de la realidad (es decir, el realismo perceptual). La imagen estereoscópica puede presentar distintos tipos de distorsiones que la hagan menos natural. Por ejemplo, los objetos estereoscópicos se perciben en ocasiones como anormalmente grandes o pequeños (efecto teatro de marionetas) o parecen anormalmente delgados (efecto papel de cartón).

Sentido de presencia, que se refiere a la experiencia subjetiva de estar en un lugar o entorno dado aunque uno se encuentra situado en otro distinto.

En esta Recomendación se presenta información relativa a los métodos y procedimientos para la evaluación de las tres dimensiones primarias señaladas: calidad de la imagen, calidad de la profundidad y confort visual. La presente Recomendación no incluye metodologías para la evaluación de la naturalidad ni del sentido de presencia, aunque su inclusión está prevista en una versión posterior.

A7-2 Metodologías subjetivas

La presente Recomendación comprende varias metodologías para la evaluación de la calidad de la imagen. En todos los métodos se realizan un conjunto de pruebas de evaluación con un grupo de observadores utilizando una serie de secuencias de vídeo, que han sido procesadas por los sistemas en estudio (por ejemplo, un algoritmo con distintos parámetros, una tecnología de codificación que admite distintas velocidades binarias, diferentes casos de transmisión etc.). En cada prueba se pide a los observadores que evalúen una característica relevante (por ejemplo, la calidad de la imagen) de la secuencia o secuencias de vídeo utilizando una escala prestablecida. Los métodos difieren entre sí principalmente en el modo de presentación, es decir, la forma en la que se presentan al observador las secuencias de vídeo y en la escala utilizada por los observadores para evaluar dichas secuencias.

Las imágenes de prueba son imágenes estéreo binoculares seleccionadas en base a los elementos descritos en § A7-4. Los observadores evalúan las tres características siguientes:

– calidad de la imagen: efecto sobre la resolución de las imágenes 3D estereoscópicas de un sistema que incluye un trayecto entre las imágenes en prueba y la pantalla donde se muestran;
– calidad de la profundidad: efecto sobre la profundidad de percepción de las imágenes 3D estereoscópicas de un sistema que incluye un trayecto entre las imágenes en prueba y la pantalla donde éstas se muestran;
– confort visual: efecto sobre la facilidad de visualización de las imágenes 3D estereoscópicas de un sistema que incluye un trayecto entre las imágenes en prueba y la pantalla donde éstas se muestran.

En este Anexo se especifican los seis métodos previstos en la presente Recomendación. Dichos métodos se han utilizado satisfactoriamente en las últimas dos décadas en trabajos de investigación sobre calidad de imagen, calidad de profundidad y confort visual de tecnologías de imágenes estereoscópicas. Los métodos en cuestión son los siguientes:

– método de estímulo único (SS, *single-stimulus*);
– método de escala de degradación con doble estímulo (DSIS, *double-stimulus impairment scale*);
– método de escala de calidad continua de doble estímulo (DSCQS, *double stimulus continuous quality scale*);

– método de comparación de estímulos (SC, *stimulus-comparison*);

– método de evaluación de calidad continua de estímulo único (SSCQE, *single stimulus continuous quality evaluation*);

– método de doble estímulo simultáneo para evaluación continua (SDSCE, *simultaneous double stimulus for continuous evaluation*).

Cuando ha sido adecuado, los métodos se han utilizado de forma ligeramente distinta, por ejemplo, con diferentes escalas de confort visual. En los Cuadros 3-10, 3-11 y 3-12 se resumen el modo de presentación y las escalas asociadas con los métodos de evaluación de la calidad de imagen, calidad de profundidad y confort visual, respectivamente.

En el punto siguiente se hace una breve descripción de cada metodología. Los elementos metodológicos comunes a todos los métodos se presentan en puntos posteriores.

CUADRO 3-10

Métodos subjetivos de evaluación de la calidad de la imagen

Modo de presentación	Duración de la secuencia	Escala binaria	Escala discreta	Escala continua
Método de estímulo único (SS) descrito en el § 6.1 del Anexo 1	~10 s		5 Excelente 4 Bueno 3 Aceptable 2 Mediocre 1 Malo	Excelente Bueno Aceptable Mediocre Malo
Método de escala de degradación con doble estímulo (DSIS) descrito en el § 4 del Anexo 1			5 Imperceptible 4 Perceptible pero no molesto 3 Ligeramente molesto 2 Molesto 1 Muy molesto	
Método de escala de calidad continua de doble estímulo (DSCQS) descrito en el § 5 del Anexo 1	~10 s			Excelente Bueno Aceptable Mediocre Malo
Método de comparación de estímulos (SC) descrito en el § 6.2 del Anexo 1	~10 s	A vs. B	−3 Mucho peor −2 Peor −1 Ligeramente peor 0 Igual 1 Ligeramente mejor 2 Mejor 3 Mucho mejor	

CUADRO 3-10 (*fin*)

Modo de presentación	Duración de la secuencia	Escala binaria	Escala discreta	Escala continua
Método de evaluación de calidad continua de estímulo único (SSCQE) descrito en el § 6.3 del Anexo 1	~3-5 min			Excelente Bueno Aceptable Mediocre Malo
Método de doble estímulo simultáneo para evaluación continua (SDSCE) descrito en el § 6.4 del Anexo 1				Fidelidad es perfecta (codificada 100) Fidelidad es nula (codificada 0)

CUADRO 3-11

Métodos subjetivos de evaluación de la calidad de la profundidad

Modo de presentación	Duración de la secuencia	Escala binaria	Escala discreta	Escala continua
Método de estímulo único (SS) descrito en el § 6.1 del Anexo 1	~10 s		5 Excelente 4 Bueno 3 Aceptable 2 Mediocre 1 Malo	Excelente Bueno Aceptable Mediocre Malo
Método de escala de degradación con doble estímulo (DSIS) descrito en el § 4 del Anexo 1			5 Imperceptible 4 Perceptible pero no molesto 3 Ligeramente molesto 2 Molesto 1 Muy molesto	
Método de escala de calidad continua de doble estímulo (DSCQS) descrito en el § 5 del Anexo 1	~10 s			Excelente Bueno Aceptable Mediocre Malo
Método de comparación de estímulos (SC) descrito en el § 6.2 del Anexo 1	~10 s	A vs. B	−3 Mucho peor −2 Peor −1 Ligeramente peor 0 Igual 1 Ligeramente mejor 2 Mejor 3 Mucho mejor	

CUADRO 3-11 (*fin*)

Modo de presentación	Duración de la secuencia	Escala binaria	Escala discreta	Escala continua
Método de evaluación de calidad continua de estímulo único (SSCQE) descrito en el § 6.3 del Anexo 1	~3-5 min			Excelente Bueno Aceptable Mediocre Malo
Método de doble estímulo simultáneo para evaluación continua (SDSCE) descrito en el § 6.4 del Anexo 1				Fidelidad es perfecta (codificada 100) Fidelidad es nula (codificada 0)

CUADRO 3-12

Métodos subjetivos de evaluación del confort visual

Modo de presentación	Duración de la secuencia	Escala binaria	Escala discreta	Escala continua
Método de estímulo único (SS) descrito en el § 6.1 del Anexo 1	~10 s		5 Muy cómodo 4 Cómodo 3 Ligeramente incómodo 2 Incómodo 1 Muy incómodo	Muy cómodo Cómodo Ligeramente incómodo Incómodo Muy incómodo
Método de escala de degradación con doble estímulo (DSIS) descrito en el § 4 del Anexo 1			5 Imperceptible 4 Perceptible pero no molesto 3 Ligeramente molesto 2 Molesto 1 Muy molesto	
Método de escala de calidad continua de doble estímulo (DSCQS) descrito en el § 5 del Anexo 1	~10 s			Muy cómodo Cómodo Ligeramente incómodo Incómodo Muy incómodo
Método de comparación de estímulos (SC) descrito en el § 6.2 del Anexo 1	~10 s	A vs. B	−3 Mucho peor −2 Peor −1 Ligeramente peor 0 Igual 1 Ligeramente mejor 2 Mejor 3 Mucho mejor	

CUADRO 3-12 (*fin*)

Modo de presentación	Duración de la secuencia	Escala binaria	Escala discreta	Escala continua
Método de evaluación de calidad continua de estímulo único (SSCQE) descrito en el § 6.3 del Anexo 1	~3-5 min			Muy cómodo Cómodo Ligeramente incómodo Incómodo Muy incómodo
Método de doble estímulo simultáneo para evaluación continua (SDSCE) descrito en el § 6.4 del Anexo 1				Fidelidad es perfecta (codificada 100) Fidelidad es nula (codificada 0)

A7-3 Condiciones generales de la observación

Las condiciones de observación (incluida la luminancia de la pantalla, la iluminación de fondo, la distancia de observación etc.) deben ser consistentes con las utilizadas en 2D, tal como se describe en el § 2.1 de la Parte 1. El motivo de dicha consistencia es doble. Por una parte, en la práctica los usuarios verán las imágenes de TV 3D en las mismas condiciones de representación y visualización que las imágenes de 2D. En segundo lugar, los avances en términos de calidad de funcionamiento de las tecnologías de vídeo TV 3D a menudo deberán medirse en relación con (es decir, «comparadas con») el avance de las normas de las tecnologías de vídeo de TVAD.

En el § 2.1 de la Parte 1 se especifican dos posibles criterios para la selección de la distancia de visualización. Debe seleccionarse la distancia de visualización nominal (DVD, *design viewing distance*). La DVD de un sistema digital es la distancia a la que dos píxeles adyacentes presentan, para el ojo del observador, un ángulo subtendido de 1 arc-min.

Cuando dos píxeles adyacentes subtienden un ángulo de 1 arc-min desde la perspectiva del ojo del observador, la menor disparidad angular (retinal) que puede ser representada por el sistema a la distancia de observación nominal (es decir, la resolución de profundidad del sistema) es igual a 1 arc-min (o su equivalente 60 arc-s). Las investigaciones han mostrado que casi el 97% de la población puede distinguir disparidades horizontales iguales o menores a 140 arc-s y que al menos el 80% puede detectar disparidades horizontales de 30 arc-s. Por tanto, la mayoría de los observadores no tendrán dificultades para resolver la disparidad más pequeña representable en los sistemas de vídeo 3D actuales a la distancia de visualización nominal.

A7-4 Material de prueba

La selección del material de prueba debe estar asociada a la cuestión experimental que pretende abordarse en el correspondiente estudio. En general, el contenido de las secuencias de prueba (deporte, películas, etc.) y sus características espacio-temporales deben ser representativas de los programas distribuidos por el servicio analizado.

Además, el contenido de las secuencias estereoscópicas también debería ser, por lo general, de visualización cómoda. El confort de visualización de imágenes estereoscópicas depende de manera crítica de las disparidades de la imagen (paralaje) y de las condiciones de visualización. En consecuencia, deben tomarse las precauciones necesarias para garantizar que las disparidades no excedan los límites señalados en el punto siguiente, salvo que el estudio tenga por objetivo específicamente la medición del confort visual. Además, cuando sea posible, se medirán e informará

de las estadísticas: media, desviación típica y distancia (mín/máx) de la distribución de la disparidad de la secuencia de prueba.

El paralaje, las inconsistencias entre las imágenes izquierda y derecha y la distribución y cambios del paralaje, pueden ser elementos que considerar en la selección de imágenes de prueba de fácil visualización estereoscópica. La relación entre las imágenes estereoscópicas 3D de fácil visualización y el paralaje, las inconsistencias entre las imágenes izquierda y derecha y la distribución y cambios del paralaje se describen en los puntos siguientes.

A7-4.1 Utilización de material de vídeo de referencia

Los investigadores pueden tener interés en incluir una secuencia de referencia en el conjunto de secuencias de prueba, en caso de estar disponible. La referencia es normalmente una versión de la secuencia de prueba sin procesamiento alguno (es decir, es la secuencia fuente original). Para los estudios estereoscópicos, la principal referencia es la secuencia estereoscópica original no procesada. Sin embargo, el plan del experimento puede incluir también como referencia la versión monoscópica (es decir, sólo una de las vistas de la secuencia fuente original); así, por ejemplo, en los estudios del confort visual puede ser de utilidad como referencia el confort visual de la referencia monoscópica. La versión monoscópica de la referencia debe presentarse en modo 3D (por ejemplo, presentar la visión izquierda a ambos ojos con la misma configuración del *hardware* 3D que para la secuencia estereoscópica real). La inclusión de la referencia en el experimento proporciona dos ventajas importantes. En primer lugar, ofrece la oportunidad de medir la transparencia (también denominada fidelidad) del algoritmo o tecnología investigada[8]. En segundo lugar, la inclusión de la referencia supone disponer de un anclaje de alta calidad que puede ayudar a estabilizar las apreciaciones[9].

A7-4.2 Límites del confort visual

Un paralaje o disparidad excesiva causa incomodidad visual, probablemente porque empeora el conflicto entre acomodación y vergencia (movimiento binocular en el cual ambos ojos se desplazan en direcciones opuestas). Por tanto, se ha sugerido que para minimizar el conflicto entre acomodación y vergencia, las disparidades en la imagen estereoscópica deben ser suficientemente pequeñas de forma que las profundidades percibidas de los objetos se encuentren en una «zona de confort». Se han propuesto varias formas de definir dichos límites. Un enfoque utiliza la medición del paralaje de la pantalla, expresado como un porcentaje del tamaño horizontal de la misma, para especificar los límites de la visualización cómoda. Se han sugerido valores del 1% para las disparidades cruzadas/negativas y del 2% para disparidades no cruzadas/positivas (para un valor total de aproximadamente un 3%). Según otro posible enfoque, la zona de confort está delimitada por la profundidad de campo del ojo. Para condiciones típicas de visualización de televisión (radiodifusión), los investigadores han supuesto una profundidad de campo de ±0,2D (dioptrías) y ±0,3D (dioptrías). Para un sistema de TVAD con una resolución de 1920 × 1080 (Recomendación UIT-R BT.709) observado desde la distancia de visualización nominal de 3,1H, dichos valores corresponden a aproximadamente ±2% y ±3% del paralaje de la pantalla. Finalmente, un tercer enfoque especifica los límites del confort en términos de disparidad retinal y fija dichos límites en ±1° del ángulo visual de las disparidades positiva y negativa.

---

[8] La transparencia (fidelidad) es un concepto que describe la calidad de funcionamiento de un códec o de un sistema en relación con un sistema de transmisión ideal sin degradación. Es fácil ver que la transparencia puede medirse comparando las apreciaciones asignadas a la secuencia de referencia con las asignadas a la secuencia procesada con el algoritmo o la tecnología investigada.

[9] Se reconoce que la estabilidad de las apreciaciones en el espacio (es decir, realizadas en distintos laboratorios) y en el tiempo (es decir, realizadas en el mismo laboratorio pero en momentos diferentes) también puede mejorarse utilizando anclajes de baja calidad. No obstante, la UIT tiene planes para definir de forma inminente anclajes de baja calidad para la evaluación de tecnologías de imágenes estereoscópicas.

Es interesante señalar que los distintos enfoques convergen en los mismos límites de confort. Recuérdese que a la distancia de visualización nominal, dos píxeles adyacentes subtienden un ángulo de 1 arc-min desde el ojo de observador. Por tanto, 60 píxeles corresponden a un ángulo visual de 1°. Ello permite especificar fácilmente los límites de confort en términos de disparidad retinal (para un observador medio). Por ejemplo, para sistemas TVAV de resolución 1920 × 1080 (Recomendación UIT-R BT.709), 1% (~19.2 píxeles) corresponde a aproximadamente 20 arc-min, 2% a ~40 arc-min y 3% a ~60 arc-min (equivalente a 1°).

Debe observarse que aunque a la distancia de visualización nominal dos píxeles adyacentes siempre subtienden un ángulo de 1 arc-min, la separación física (por ejemplo, en mm) entre dichos píxeles aumenta para pantallas más grandes (el número de píxeles es el mismo, pero aumenta el tamaño de la pantalla). Por tanto, los límites más altos (por ejemplo, ±3%) pueden hacer que para las pantallas de mayor tamaño la distancia física entre puntos correspondientes (es decir, el paralaje de dos vistas en mm) sea superior a la distancia entre pupilas de un observador medio (~63-65 mm). Ello puede dar lugar a una creciente incomodidad.

A7-4.3 Discrepancias entre las imágenes izquierda y derecha

En sistemas 3D estereoscópicos, la imagen binocular se forma presentando la imagen izquierda y derecha a los respectivos ojos. Si existen discrepancias entre ambas imágenes, puede producirse estrés psicofísico y, en algunos casos, puede no ser posible la visualización 3D. Por ejemplo, cuando se filman y se muestran en pantalla programas estereoscópicos de TV 3D, pueden producirse entre las imágenes izquierda y derecha distorsiones geométricas tales como inconsistencia del tamaño, desplazamiento vertical o error de rotación. Es conveniente que las imágenes de prueba estén libres de dichas distorsiones geométricas. Para más información véase el § 3.2.1 del Anexo 4 al Informe UIT-R BT.2160-2.

Los elementos relativos a discrepancias entre las imágenes izquierda y derecha que deben considerarse cuando se seleccionen imágenes estereoscópicas TV 3D de prueba que sean de fácil visualización son los siguientes:

– discrepancia geométrica, incluyendo el tamaño, desplazamiento vertical y rotación;
– discrepancia en el brillo, incluyendo los niveles de blanco y de negro;
– diafonía.

A7-4.4 Distancia, distribución y cambio en el paralaje

Las distribuciones del paralaje están correladas con el confort visual de las imágenes estereoscópicas.

La distribución de paralaje de las imágenes estereoscópicas es discontinua durante los cuadros de cambio de escena. Cuando se produce un paralaje extremo o cambios bruscos en el paralaje se produce un efecto visual incómodo, por lo que es importante mantener un nivel adecuado de paralaje de las imágenes. Para más información véase el § 3.2.2 del Anexo 4 al Informe UIT-R BT.2160-2.

Por lo general, dado que los estudios que utilizan secuencias de prueba estereoscópicas pueden generar un cierto grado de incomodidad visual, es recomendable utilizar, siempre que sea posible, material de prueba cuyas disparidades no excedan los límites de confort, aunque ocasionalmente puedan permitirse situaciones puntuales en las que se superen dichos límites.

A7-5 Equipamiento para la experimentación

Los aparatos utilizados en los experimentos (servidor de vídeo, pantalla, etc.) deben permitir visualizar secuencias de prueba 3D con la máxima resolución, por ejemplo, utilizando un formato de empaquetamiento de trama HDMI. Ello ofrece una mayor flexibilidad respecto al conjunto de estudios que pueden realizarse.

Hasta la fecha, no se ha normalizado una pantalla TV 3D de referencia. En consecuencia, es previsible que la mayoría de los investigadores utilicen pantallas TV 3D de gran consumo. Dado que las características de dichas pantallas pueden variar de un fabricante a otro, se insta a los investigadores a que informen de los ajustes relevantes de la pantalla utilizados en cada estudio.

A7-6 Observadores

A7-6.1 Tamaño de la muestra

En general, es recomendable utilizar al menos 30 observadores. No obstante, el número real dependerá de los objetivos específicos de la investigación, y teniendo en cuenta que las consideraciones sobre el tamaño del grupo para estudios de TV 3D no difieren de los de 2D.

A7-6.2 Examen de la visión de los observadores

Los observadores deberán pasar un examen de agudeza visual, daltonismo y visión estereoscópica aplicando las pruebas clínicas vigentes, tales como las cartas de Snellen para la agudeza visual; los diagramas de Ishara, o equivalente, para el color, y la prueba de Randot, o equivalente, para la visión estereoscópica. Obsérvese que las pruebas de visión estereoscópica como la de Randot, la de la mosca o la de Frisby, miden, por lo general, disparidades retinales de entre 20 y 400 arc-s. Se insta a que los investigadores informen de las estadísticas relevantes de las capacidades estereoscópicas de los observadores participantes en un estudio. Si fuera necesario un análisis más detallado de dichas capacidades estereoscópicas, los investigadores pueden utilizar los materiales de prueba que se incluyen en el Adjunto 1 al presente Anexo.

A7-7 Instrucciones a los observadores

Las instrucciones a los observadores deben adaptarse a las dimensiones investigadas (por ejemplo, calidad de profundidad, confort, etc.). En particular, las directrices que es necesario dar a los observadores para la realización de estudios sobre imágenes en 3D deben ser más estrictas que las típicamente utilizadas en la evaluación de la calidad de imágenes en 2D, ya que los participantes pueden experimentar malestar visual. En general, los estudios sobre 3D exigen dar mayor grado de información a los participantes sobre las razones del estudio, así como los posibles efectos negativos que pudiera producir la exposición a los estímulos.

A7-8 Duración de la sesión

Si el material de visualización se considera confortable, la prueba podría tener una duración tan larga como la de los estudios de 2D (es decir, ~20-40 minutos con interrupciones entre ellas). Si el material de prueba tiene un paralaje excesivo, y por tanto es potencialmente molesto, debe limitarse la duración.

A7-9 Variabilidad de las respuestas

Las apreciaciones de observadores en experimentos de evaluación subjetiva son, por lo general, bastante variables. Las diferencias entre observadores puede que simplemente reflejen las características de la población de referencia y, por tanto, pueden resolverse aumentando el tamaño de la muestra.

Sin embargo, parte de la variabilidad puede deberse a cambios en los patrones de respuesta de los observadores individuales durante el experimento. Dichos cambios implican una modificación de los criterios de evaluación debidos, por ejemplo, a la mayor práctica en la realización de la prueba, al aprendizaje de las características de los efectos molestos, etc.). Para minimizar los efectos negativos de dicha variabilidad, los investigadores deben establecer procedimientos de entrenamiento adecuados (tarea, nivel de degradación, etc.), utilizar la aleatorización múltiple (es decir, presentar

las secuencias de prueba en órdenes aleatorios diferentes a los distintos observadores) y replicar las secuencias (lo que, además, permitiría medir los posibles cambios en los patrones de respuesta).

A7-10 Criterios de rechazo de observadores

Los criterios para el rechazo de observadores (selección de los observadores) correspondientes a los métodos identificados en § A7-2 se describen en la Parte 1.

A7-11 Análisis estadístico

Los análisis estadísticos para la investigación de sistemas de imágenes en 3D son los mismos que para sistemas de imágenes en 2D.

Adjunto 1
al Anexo 7

Materiales de prueba para las pruebas de visión

A7-1 Prueba de visión

En el Cuadro 3-13 se enumeran los diagramas utilizados en las pruebas de visión. Las 12 pruebas se han seleccionado de conformidad con la jerarquía del sistema de visión humano, desde el nivel más bajo al más alto. A continuación se describen ocho pruebas de visión (VT, *visión test*) principales, siendo las cuatro restantes utilizadas en pruebas clínicas. Los observadores deben tener un estereopsis normal, es decir, deben pasar la prueba VT-04 de la estereopsis fina y la VT-07 para la estereopsis dinámica. Las restantes seis pruebas se utilizan para una caracterización más detallada. Los diagramas de prueba deben visualizarse a una distancia de tres veces la altura de la pantalla en las que se proyectan.

CUADRO 3-13

Materiales de prueba estereoscópica para una prueba de visión

N°	Elemento	Prueba de	Contenido
1	Recepción simultánea	Capacidad de percibir simultáneamente imágenes presentadas dicópticamente y en la posición correcta	Se presenta la imagen de una jaula en un ojo y la de un león en el otro
2	Fusión binocular	Capacidad de percibir dos imágenes dicópticas en los ojos izquierdo y derecho como una sola imagen	La imagen para un ojo tiene dos puntos y la del otro ojo tiene tres puntos, con un punto común
3	Estereopsis gruesa	Capacidad de percibir imágenes que se presentan de forma dicóptica con un cierto paralaje como una sola imagen con una profundidad considerable	Las imágenes para ambos ojos son una estereopareja de imágenes de una libélula con sus alas extendidas
4	Estereopsis fina	Capacidad de percibir imágenes que se presentan de forma dicóptica con un cierto paralaje como una sola imagen con una profundidad reducida	Se presentan nueve romboides de prueba, cada uno de los cuales tiene cuatro círculos, y sólo uno de los círculos tiene un pequeño paralaje

CUADRO 3-13 (*fin*)

N°	Elemento	Prueba de	Contenido
5	Límite de fusión cruzada	Capacidad de percibir imágenes que se presentan de forma dicóptica con disparidades cruzadas como una sola imagen	Se presenta una estereopareja de barras cuyo paralaje varía a razón de 10'/s
6	Límite de fusión sin cruce	Capacidad de percibir imágenes presentadas de forma dicóptica con disparidades no cruzadas como una sola imagen	Se presenta una estereopareja de barras cuyo paralaje varía a razón de 11'/s
7	Estereopsis dinámica	Capacidad de percibir la profundidad en imágenes de un estereograma de puntos aleatorios en movimiento	Estereograma de puntos aleatorios dinámicos
8	Agudeza binocular	Agudeza binocular, incluyendo cualquier asimetría de la agudeza monocular que pueda impedir una buena estereopsis	Caracteres E con diversas orientaciones y tamaños
9	Estrabismo horizontal	Desviación horizontal del ojo que el paciente no puede evitar	Líneas verticales y horizontales
10	Estrabismo vertical	Desviación vertical del ojo que el paciente no puede evitar	Líneas verticales y horizontales
11	Aniseiconia	Condición en la que la imagen ocular de un objeto visto por un ojo difiere en tamaño y forma respecto a como lo ve el otro ojo	La imagen izquierda consiste en caracteres «[o» y la derecha consiste en caracteres «o]», donde el carácter «o» tiene la misma posición en ambas
12	Cicloforia	Desviación de uno de los ojos alrededor del eje anteroposterior cuando se evita la fusión	La imagen izquierda consiste en la superficie de un reloj y la derecha en las manecillas del reloj marcando las seis en punto

NOTA 1 – Estos materiales están en el formato 1125/60/I (véase la Recomendación UIT-R BT.709).

NOTA 2 – Estos materiales pueden obtenerse del Institute of Image Information and Television Engineers (ITE), 3-5-8 Shibakoen, Minato-ku, Tokio 105-0011, Japón. Tel.: 81-3-3432-4677, correo electrónico: ite@ite.or.jp.

Las imágenes en miniatura siguientes, a derecha e izquierda, se colocan una junto a otra con fines explicativos para una fusión sin cruces.

1) *VT-01*: Percepción simultánea (prueba del león)

Prueba la capacidad de percibir simultáneamente imágenes presentadas dicópticamente y en la posición correcta. Se presenta la imagen de una jaula en un ojo y la de un león en el otro, cuya posición se desplaza a razón de 12'/s. El tamaño de cada imagen se fija a 10°, de tal forma que los observadores pueden capturar las imágenes en sus paramáculas. Los observadores con una visión normal pueden ver al león dentro de la jaula durante un cierto tiempo del periodo de presentación.

FIGURA 3-8

Diagrama de prueba para VT-01

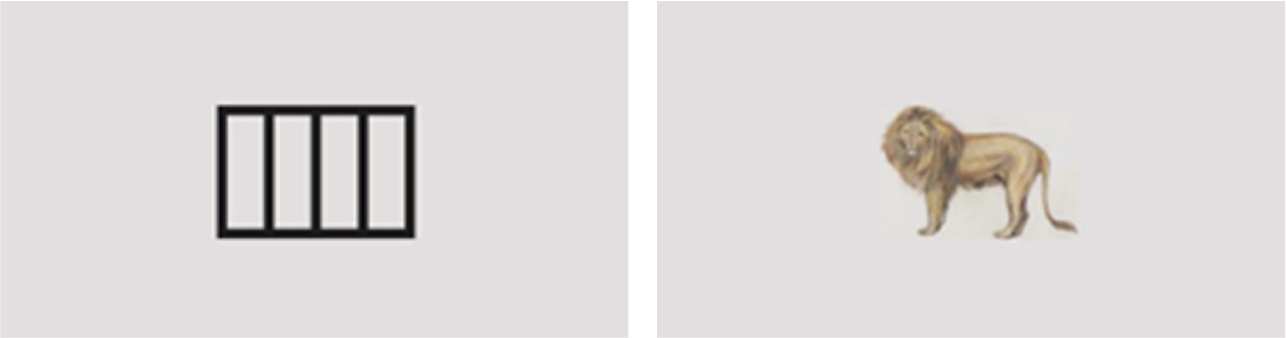


Imagen derecha Imagen izquierda

BT.0500-03-08

2) *VT-02*: Fusión binocular (prueba de los 4 puntos de Worth)

Prueba de la capacidad de percibir dos imágenes dicópticas en los ojos izquierdo y derecho como una sola imagen. La imagen para un ojo tiene dos puntos y la del otro ojo tiene tres puntos, con un punto común. Los observadores con una visión normal ven cuatro puntos.

FIGURA 3-9

Diagrama de prueba para VT-02

Imagen derecha Imagen izquierda

BT.500-03-09

3) *VT-03*: Estereopsis gruesa (prueba de la libélula)

Prueba de la capacidad de percibir imágenes que se presentan de forma dicóptica con un cierto paralaje como una sola imagen con una profundidad considerable. Las imágenes para ambos ojos son una estereopareja de imágenes de una libélula con sus alas extendidas. Los observadores con una visión normal perciben las alas delante de la pantalla de visualización.

FIGURA 3-10

Diagrama de prueba para VT-03

Imagen derecha Imagen izquierda

BT.0500-03-10

4) *VT-04*: Estereopsis de detalle (prueba del círculo)

Prueba la capacidad de percibir imágenes que se presentan de forma dicóptica con un cierto paralaje como una sola imagen con una profundidad reducida. Se presentan nueve romboides de prueba, cada uno de los cuales tiene cuatro círculos, y sólo uno de los círculos tiene un pequeño paralaje. Los observadores con visión normal pueden percibir el círculo con el pequeño paralaje delante de la pantalla de visualización. El Cuadro 3-14 muestra el número de prueba, las respuestas correctas y el ángulo de estereopsis a 3 *H*.

CUADRO 3-14

Respuestas correctas y paralaje

Número de prueba	**Repuesta correcta**	**Ángulo de estereopsis a 3 *H* (")**
1	Abajo	480
2	Izquierda	420
3	Abajo	360
4	Arriba	300
5	Arriba	240
6	Izquierda	180
7	Derecha	120
8	Izquierda	60
9	–	0

FIGURA 3-11

Diagrama de prueba para VT-04

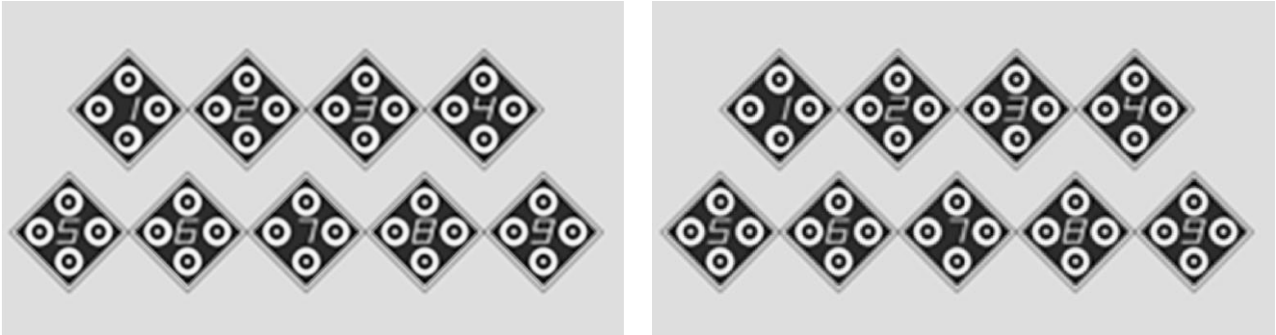


Imagen derecha Imagen izquierda

BT.0500-03-11

5) *VT-05*: Límite de fusión cruzada (prueba de la barra)

Prueba la capacidad de percibir imágenes que se presentan de forma dicóptica con disparidades cruzadas como una sola imagen. Se presenta una estereopareja de barras cuyo paralaje varía a razón de 10'/s. Pueden medirse los límites de fusión de las series ascendentes y descendentes. Se pide a los observadores que informen del momento en que detectan la ruptura de fusión, es decir, tan pronto como perciben imágenes dobles en las series ascendentes, así como de la recuperación de la fusión, es decir, tan pronto como perciben las imágenes dicópticas como una imagen única en las series descendentes.

FIGURA 3-12

Diagrama de prueba para VT-05

Imagen derecha Imagen izquierda

BT.0500-03-12

6) *VT-06*: Límite de fusión sin cruce (prueba de la barra)

Prueba la capacidad de percibir imágenes presentadas de forma dicóptica con disparidades no cruzadas como una sola imagen. Las imágenes que se presentan son las mismas que en el caso cruzado anterior, pero se invierten las imágenes derecha e izquierda.

FIGURA 3-13

Diagrama de prueba para VT-06

Imagen derecha Imagen izquierda

BT.0500-03-13

7) *VT-07*: Estereopsis dinámica (prueba del estereograma de puntos aleatorios dinámicos)

Prueba la capacidad de percibir la profundidad en imágenes de un estereograma de puntos aleatorios en movimiento. Los observadores con visión normal pueden percibir una forma rectangular y un movimiento sinusoidal en profundidad en el estereograma de puntos aleatorios dinámicos.

FIGURA 3-14

Diagrama de prueba para VT-07

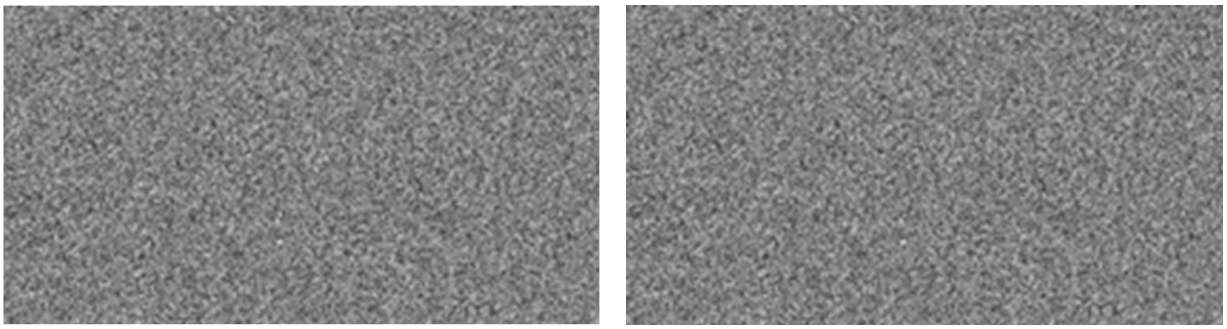


Imagen derecha Imagen izquierda

BT.0500-03-14

8) *VT-08*: Agudeza binocular (prueba de agudeza)

Prueba la agudeza binocular con fusión binocular, incluyendo cualquier asimetría de la agudeza monocular que pueda impedir una estereopsis adecuada. Las imágenes tienen cuatro columnas y cinco líneas que consisten en caracteres E con diversas orientaciones y tamaños. Las dos columnas centrales pueden verse con ambos ojos; las dos columnas de la izquierda sólo pueden verse con el ojo izquierdo y las dos columnas de la derecha sólo pueden verse con el ojo derecho. Los observadores con una visión normal pueden decir correctamente la orientación de los caracteres E. Los tamaños de los caracteres se corresponden con agudezas de 1,0, 0,5, 0,33, 0,25 y 0,125 a 3 *H*.

FIGURA 3-15

Diagrama de prueba para VT-08

Imagen derecha Imagen izquierda

BT.0500-03-15

9 y 10) *VT-09*: Estrabismo horizontal (prueba horizontal de Maddox) y *VT-10*: Estrabismo vertical (prueba vertical de Maddox)

Estos diagramas miden la desviación horizontal y vertical del ojo. Los ejes visuales tienen una posición relativa entre ellos distinta, según las condiciones fisiológicas. Las imágenes constan de una línea vertical y otra horizontal. Los observadores con una visión normal perciben el punto de cruce de las líneas, aproximadamente, en el centro de las mismas. Los números junto a las marcas indican las dioptrías prismáticas para una distancia pupilar (PD) de 65 mm a 3,02 *H*.

FIGURA 3-16

Diagrama de prueba para VT-09

BT.0500-03-16

FIGURA 3-17

Diagrama de prueba para VT-10

BT.0500-03-17

11) *VT-11*: Aniseiconia («[ ]»prueba de caracteres)

Condición en la que la imagen ocular de un objeto visto por un ojo difiere en forma y tamaño de la vista por el otro ojo. La imagen izquierda consiste en caracteres «[o» y la imagen derecha consiste en caracteres «o]», donde el carácter «o» tiene la misma posición en ambas. Los observadores con una visión normal perciben los caracteres «[» y «]» con el mismo tamaño y la misma altura.

FIGURA 3-18

Diagrama de prueba para VT-11

BT.0500-03-18

12) *VT-12*: Cyclophoria (prueba del reloj)

Desviación de uno de los ojos alrededor del eje anteroposterior cuando se evita la fusión. La imagen izquierda consiste en la superficie de un reloj y la imagen derecha en las manecillas del reloj marcando las seis en punto.

FIGURA 3-19

Diagrama de prueba para VT-12

BT.0500-03-19